

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/88732>

**Copyright and reuse:**

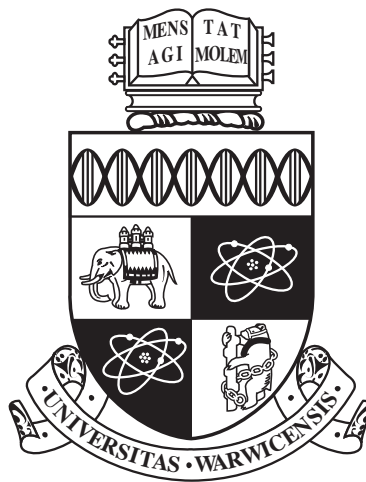
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)



**Spatio-temporal Framework on Facial expression  
Recognition**

by

**Xijian Fan**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**School of Engineering**

July 2016

THE UNIVERSITY OF  
**WARWICK**

# Contents

<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>x</b>
<b>Acknowledgments</b>	<b>ii</b>
<b>Declarations</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Publications</b>	<b>vi</b>
<b>Abbreviations</b>	<b>vii</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Background knowledge . . . . .	1
1.1.1 Perspectives from psychology . . . . .	2
1.2 Objectives and motivation . . . . .	7
1.3 Contributions . . . . .	8
1.4 Structure of thesis . . . . .	9
<b>Chapter 2 Literature review on Facial expression recognition</b>	<b>10</b>
2.1 Techniques in facial expression recognition systems . . . . .	10
2.1.1 Face acquisition . . . . .	11
2.1.2 Feature representation . . . . .	17
2.1.3 Classification . . . . .	28
2.1.4 Challenges . . . . .	30
2.2 Datasets for facial expression recognition . . . . .	30
2.3 Multi-modal expression analysis . . . . .	32
2.4 Problem Space for facial expression analysis . . . . .	33

2.4.1	Posed vs. spontaneous expression . . . . .	34
2.4.2	Under-control condition vs. real life . . . . .	34
2.4.3	Individual difference in subjects . . . . .	35
2.4.4	Transitions among expressions . . . . .	35
<b>Chapter 3 Patch based facial expression recognition framework</b>		<b>37</b>
3.1	Introduction . . . . .	37
3.2	Proposed framework . . . . .	38
3.3	Facial landmark detection . . . . .	39
3.3.1	Pre-processing . . . . .	40
3.3.2	Eye and Nose Localisation . . . . .	40
3.3.3	Lip corner detection . . . . .	41
3.3.4	Eyebrow corner detection . . . . .	43
3.4	Extraction of active facial patches . . . . .	44
3.5	Feature extraction and classification . . . . .	45
3.5.1	Sparse representation . . . . .	45
3.5.2	Learning salient facial patches across expressions . . . . .	47
3.6	Experiments . . . . .	49
3.6.1	Experiments on the Cohn-Kanade database . . . . .	50
3.6.2	Experiments on JAFFE Database . . . . .	51
3.6.3	Experiments on fused database . . . . .	51
3.7	Conclusion . . . . .	52
<b>Chapter 4 Spatial-Temporal Framework Based on Histogram of Gradient and Optical Flow</b>		<b>54</b>
4.1	Introduction . . . . .	54
4.2	Proposed framework . . . . .	55
4.3	Dynamic features extraction . . . . .	58
4.3.1	PHOG_TOP descriptor . . . . .	58
4.3.2	Dense optical flow descriptor . . . . .	60
4.3.3	Integration of descriptors . . . . .	63
4.4	Experiments . . . . .	64
4.4.1	Facial expression datasets . . . . .	64
4.4.2	Experimental results . . . . .	65
4.5	Conclusion . . . . .	74

<b>Chapter 5</b>	<b>Spatio-temporal Framework Based on Local Zernike Mo-</b>	
	<b>ment and Motion History Image</b>	<b>75</b>
5.1	Introduction . . . . .	75
5.2	Proposed framework . . . . .	77
5.2.1	Dimensionality reduction using 2D PCA . . . . .	77
5.3	Feature extraction . . . . .	78
5.3.1	Motion History Image . . . . .	78
5.3.2	Optical Flow Algorithm . . . . .	79
5.3.3	Optical Flow based MHI (MHI_OF) . . . . .	80
5.3.4	Entropy . . . . .	81
5.3.5	Local Zernike Moment . . . . .	83
5.3.6	Extension to spatio-temporal . . . . .	85
5.3.7	Fusion using weighting function . . . . .	86
5.4	Experiments . . . . .	87
5.4.1	Facial expression datasets . . . . .	87
5.4.2	Experimental results . . . . .	87
5.5	Conclusion . . . . .	92
<b>Chapter 6</b>	<b>Conclusions and Future Work</b>	<b>94</b>
6.1	Conclusions . . . . .	94
6.2	Future work . . . . .	96

# List of Tables

2.1	Number of image sequences (subjects) for each expression in the CK+ dataset. . . . .	32
3.1	The confusion matrix using Framework 1 on CK+ database. . . . .	50
3.2	The confusion matrix using Framework 1 on JAFFE database. . . . .	51
3.3	The salient patches derived from fusion CK+ and JAFFE Dataset. . . . .	52
4.1	Number of image sequences (subjects) for each expression in the CK+ dataset and MMI dataset. . . . .	65
4.2	Multiclass SVM results of PHOG_TOP from the whole face on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. . . . .	68
4.3	Multiclass SVM results in using Dense flow optical flow on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	69
4.4	Multiclass SVM results in using PHOG_TOP on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	70
4.5	Multiclass SVM results in using Framework 2 on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	70
4.6	Comparative evaluation of Framework 2 with 2 methods using leave-subject-out cross-validation. . . . .	71

4.7	Confusion matrix in using dense flow optical flow on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using dense optical flow and in using shape.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	72
4.8	Confusion matrix in using PHOG.TOP on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using PHOG.TOP and in using appearance (of Lucey et al.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	72
4.9	Confusion matrix in using Framework 2 on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using the combined features of dense optical flow and PHOG.TOP and in using the combined features of shape and appearance (of Lucey et al.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error. . . . .	73
4.10	Comparative evaluation of Framework 2 using MMI dataset . . . . .	74
5.1	Recognition rate of enMHI_OF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. . . . .	88
5.2	Recognition rate of QLZM.MCF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. . . . .	89
5.3	Recognition rate of enMHI_OF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. . . . .	89
5.4	Recognition rate of using simple fusion strategy on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation. . . . .	90

5.5	Recognition rate of the proposed fusion strategy on classification of six facial expressions of the CK+ dataset and contempt with leave-sequence-out cross-validation. . . . .	90
5.6	The overall recognition rates of the four spatio-temporal features on the CK+ dataset. . . . .	90
5.7	Comparative evaluation of Framework 3 on the MMI dataset. . . . .	91
5.8	Number of image sequences (subjects) for each expression in the AFEW dataset. . . . .	92
5.9	Recognition rate of the proposed strategy on classification of six basic facial expressions and neutral expression of the AFEW dataset. . . .	93



# List of Figures

1.1	Example of expression for the six basic emotions (Left-to-right from top row: anger, fear, disgust, surprise, happiness, and sadness). . . .	4
1.2	Example of AUs which defined in FACS . . . . .	5
1.3	Facial Action Units. . . . .	5
1.4	Some examples of Action Units. Action Units are atomic facial muscle actions described in FACS. . . . .	6
2.1	Basic structure of facial expression recognition system . . . . .	11
2.2	The sum of the pixels within rectangle D can be computed with four array references. The value of the integral image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is A + B, At location 3 is A + C, and at location 4 is A + B + C + D. The sum within D can be computed as $4 + 1 - (2 + 3)$ . . . . .	15
2.3	The left, middle, and right graphics above show the absolute value, and the real and imaginary components of a sample Gabor filter. . .	19
2.4	Three examples of the extended: the circular (8; 1) neighbourhood, the circular (12; 1) neighbourhood, and the circular (16; 2) neighbourhood, respectively . . . . .	21
2.5	PHOG descriptor of an image. . . . .	22
2.6	Example of expression for the six basic emotions (Left-to-right from top row: anger, disgust, fear, happiness, sadness, and surprise). . . .	31
2.7	Example of the face and body feature extraction employed in the FABO system. (a) Face features. (b) Body features shoulder extraction procedure. Shoulder regions found and marked on the neutral frame (first row), estimating the movement within the shoulder regions using optical flow (second row) . . . . .	33
3.1	Framework 1 for facial expression recognition. . . . .	39

3.2	Framework for automated facial landmark detection and active patch extraction: (a) face detection, (b) coarse ROI selection for eyes and nose, (c) eyes and nose detection followed by coarse ROI selection for eyebrows and lips, (d) detection of corners of lip and eyebrows, (e) finding the facial landmark locations, and (f) extraction of active facial patches. . . . .	40
3.3	Lip corner localisation: (a) lips ROI, (b) applying horizontal Sobel edge detector, (c) applying Otsu thresholding, (d) removing spurious edges, and (e) applying morphological operations to render final connected component for lib corner localisation. . . . .	42
3.4	Lip corner localisation where the upper lip is not entirely connected, (a-c) same as Figure 3.3, (d-e) selection of two connected components by scanning from top, and (f) localized lip corners. . . . .	42
3.5	Eyeblink corner localisation: (a) rectangles enclosing ROI and plus marks showing the detection result, (b and f) eye ROIs, (c and g) applying adaptive threshold on ROIs, (d and h) applying horizontal Sobel edge detector followed by Otsu thresholding and morphological operations, and (e and i) final connected components for corner localisation. . . . .	43
3.6	Position of facial patches. . . . .	45
3.7	(a) Sparse decomposition of a facial image including images of the same person in the dictionary; (b) Sparse decomposition of an expressive facial image including images of the same person in the dictionary; (c) Sparse decomposition of the same image excluding images of the same facial class from the dictionary; and (d) Sparse decomposition of the differences images. . . . .	48
3.8	Eyeblink corner localisation: (a) rectangles enclosing ROI and plus marks showing the detection result, (b and f) eye ROIs, (c and g) applying adaptive threshold on ROIs, (d and h) applying horizontal Sobel edge detector followed by Otsu thresholding and morphological operations, and (e and i) final connected components for corner localisation. . . . .	49
3.9	The recognition rate using different number of salient patches. . . .	50
4.1	Framework 2 for facial expression recognition. . . . .	56
4.2	PHOG descriptor of a face. . . . .	58

4.3	Three planes in spatio-temporal domain for extracting TOP features, and the histogram concatenated from three planes: (a) original image, (b) the x-y, y-t, and x-t planes, and the concatenation of resulting histograms into a single feature set. . . . .	59
4.4	(Left) four facial sub-regions, and (right) face video sequence with 3D mouth sub-region. . . . .	61
4.5	(Left) a neutral image, and (right) with the dense optical flow (denoted by needles) superimposed. The magnitude and direction of the flow are respectively indicated by its length and the direction of its arrow. . . . .	62
4.6	Example of expressions from CK+ dataset (top row) and MMI dataset (bottom row). Each image is the frame with the most expressive face in a video sequence. . . . .	64
4.7	Recognition rates of all expressions PHOG from either XY, XT, YT or their combination, i.e., PHOG_TOP. . . . .	66
4.8	Recognition rates of all expressions in using combination of non-weighted and weighted PHOG_TOP. . . . .	67
4.9	Discriminant power of of facial sub-regions in recognising six expressions using PHOG_TOP. . . . .	69
5.1	Framework 3 for facial expression recognition. . . . .	77
5.2	Example of images from sequences (left and middle) and its MHI (right). . . . .	79
5.3	Optical flow based MHI for Anger, Happiness and Surprise (from right to left). . . . .	80
5.4	Example image entropies: (left column) neutral image and its entropy; (middle colulmn) surprise image and its entropy; and (right column) MHI of surprise image and its entropy. Lighter shades denote larger entropy values. . . . .	82
5.5	QLZM based facial representation framework. . . . .	84
5.6	Recognition rates of all expressions using QLZM and QLZM_MCF. . . . .	88
5.7	Recognition rates of all expressions using MHI and MHI_OF. . . . .	89
5.8	Sample frames from AFEW dataset. . . . .	92

# Acknowledgments

I would like to thank Dr Tardi Tjahjadi for supervising me on facial expression recognition. During his supervision, I have read, learnt and implemented a number of algorithms, from which I was able to obtain ideas on what to do and how in order to fulfil my PhD requirements. Dr Tardi is a very studious and hard-working person, who arrives in the School at 7.30 am every morning and leaves at 6.00 pm. He often works extra hours in the weekends to revise my manuscripts. His conscientious personality also helped me correct numerous errors in my articles and encourage me to be more careful. Thank you very much for helping me to submit three journal articles, two of which has been published.

I would like to thank my lab-mates AngLi, Sruti, and Ting at Image Processing and Expert Systems Laboratory, for providing various kinds of help. I would like to thank my PhD friends, Zhaohui Wang, Junyi, Guannan, and Jieren for giving me constant strength for four years. I want to thank my parents and other family relatives as well.

Finally, I would like to express my gratitude to the School of Engineering of Warwick University and China Scholarship Council for providing the financial support throughout my PhD study.

# Declarations

I hereby declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other publication unless otherwise specified. The thesis work was conducted from 2012 to 2016 under the supervision of Reader Dr. Tardi Tjahjadi at Image Processing Laboratory, School of Engineering, University of Warwick.

# Abstract

This thesis presents an investigation into two topics that are important in facial expression recognition: how to employ the dynamic information from facial expression image sequences and how to efficiently extract context and other relevant information of different facial regions. This involves the development of spatio-temporal frameworks for recognising facial expression.

The thesis proposed three novel frameworks for recognising facial expression. The first framework uses sparse representation to extract features from patches of a face to improve the recognition performance, where part-based methods which are robust to image alignment are applied. In addition, the use of sparse representation reduces the dimensionality of features, and improves the semantic meaning and represents a face image more efficiently.

Since a facial expression involves a dynamic process, and the process contains information that describes a facial expression more effectively, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Thus, the second framework uses two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on PHOG (pyramid histogram of gradient) to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. The framework views an image sequence as a spatio-temporal volume, and uses temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. Specifically, spatial based descriptor representing spatial local shape is extended to spatio-temporal domain to capture the changes in local shape of facial sub-regions in the temporal dimension to give 3D facial component sub-

regions of forehead, mouth, eyebrow and nose. The descriptor of optical flow is also employed to extract the information of temporal. The fusion of these two descriptors enhance the dynamic information and achieves better performance than the individual descriptors.

The third framework also focuses on analysing the dynamics of facial expression sequences to represent spatial-temporal dynamic information (i.e., velocity). Two types of features are generated: a spatio-temporal shape representation to enhance the local spatial and dynamic information, and a dynamic appearance representation. In addition, an entropy-based method is introduced to provide spatial relationship of different parts of a face by computing the entropy value of different sub-regions of a face.

# Publications

1. X. Fan, and T. Tjahjadi. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition* 48.11 (2015): 3407-3416.
2. X. Fan, and T. Tjahjadi. A Dynamic Framework Based on Local Zernike Moment and Motion History Image for Facial Expression Recognition. (accepted by *Pattern Recognition*)
3. X. Fan, and T. Tjahjadi. Patch based facial expression recognition. (under preparation)



# Abbreviations

- 1D: 1-dimensional
- 2D: 2-dimensional
- 3D: 3-dimensional
- AFEW: acted facial expressions in the wild
- AUs: action units
- FACS: Facial Action Coding System
- FER: facial expression recognition
- GP: graph preserving
- GUI: graphical user interface
- KNN: K-Nearest Neighbour
- HMM: hidden Markov models
- HOG: histogram of gradient
- HSI: hue, saturation, and intensity
- IMU: inertial measurement unit
- LBP: local binary pattern
- LEAR: local evidence aggregated regression
- LGBP: local Gabor binary pattern

- LDA: linear discriminant analysis
- LLE: locally linear embedding
- LZM: local Zernike moment
- MEI: motion energy image
- MCF: motion change frequency
- MHI: motion history image
- NMF: non-Negative matrix factorisation
- OF: optical flow
- PCA: principal component analysis
- QLZM: quantised local Zernike moment
- RBF: radial basis function
- RGB: red, green and blue
- PHOG: pyramid histogram of gradient
- ROI: region of interest
- SD: subclass discriminant
- SRC: sparse-representation based classifier
- SVM: support vector machine
- TOP: three orthogonal plane
- VLBP: volume local binary patterns

# Chapter 1

## Introduction

This thesis presents an investigation into two topics that are important in facial expression recognition: how to employ the dynamic information from facial expression image sequences and how to efficiently and robustness extract context and other relevant information of different facial regions. This involves the development of spatio-temporal frameworks for recognising facial expression. Section 1.1 provides the introduction of background knowledge, which includes the overview of facial expression studies from the cognition science. The research commencing with the objectives and motivation is then introduced in Section 1.2. In addition, the main contributions are briefly introduced in Section 1.3. The organisation of this thesis are finally presented in Section 1.4.

### 1.1 Background knowledge

With the rapid development of the technology of computer and artificial intelligence, the demand for human and computer interaction (HCI) has been increasingly higher. Verbal and non-verbal, as two main forms of communication play a vital role in our daily life and complete various routine tasks [1; 2; 3]

For the last forty years (specifically since 1974 [4]), computer vision research community has shown a lot of interest in analysing and automatically recognising facial expressions. Initially inspired by the findings of the cognitive scientists, the computer vision/science research community envisioned to develop such frameworks that can do the job of expression recognition in videos or still images.

As we know, it is easy for people to recognise the emotion expression of others when they communicate face to face. However, it is relatively difficult for computer. For motion recognition in computer vision, higher level knowledge is required. For example, although facial expressions can convey emotion, they can

also expression intention, cognitive processes, physical effort, or other intra or interpersonal meanings. Interpretation is assisted by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing [5; 6; 7].

Facial expressions are studied simultaneously by different scientists from different domains i.e. cognitive scientists, psychologists, neuroscientists and computer scientists etc. Although the algorithms proposed in this research work fall in the category of "computer vision" but they are based on the psycho-visual study. As the psycho-visual study is based on the principles of cognitive science, Section 1.1.1 provides the overview of facial expression studies from the cognition science literature.

### 1.1.1 Perspectives from psychology

Communication in any form i.e. verbal or non-verbal is vital to complete various daily routine tasks and plays a significant role in life. Facial expression is one of the most effective form of non-verbal communication and it provides a clue about emotional state, mindset and intention [8; 9; 10; 3]. Facial expressions not only can change the flow of conversation [11] but also provides the listeners a way to communicate a wealth of information to the speaker without even uttering a single word [12]. According to [13; 14] when the facial expression does not coincide with the other communication i.e. spoken words, then the information conveyed by the face gets more weight in decoding information. We are used to see face with different expressions every day, but still sometimes we fail to understand them. It is such paradox of the obvious and the mysterious that intrigues people to study the face and facial expressions.

Facial expression can be interpreted at different levels. The most widely used facial expression description is the Facial expressions are studied since ancient times, one of the reason is that it is one of the most important channel of non-verbal communication [15]. Initially, facial expressions were studied by great philosophers and thinkers like Aristotle and Stewart. With Darwin, the study of facial expressions became an empirical study. Darwin's studies created large interest among psychologists and cognitive scientists. The 20th century saw many studies relating facial expression to emotion and inter-human communication. Most notably, Paul Ekman reinvestigated Darwin's work and claimed that there are six universal emotions(see Figure 1.1), which are produced and recognised independently of cultural background [3].

Facial expressions can either be interpreted in terms of shown affective states

(emotions) or in terms of activated facial muscles underlying the displayed facial expression. These two approaches originate directly from the two major approaches of facial expression measurement in psychological research: message and sign judgement [16]. Message based approaches infer to what underlies a displayed facial expression, such as affect. Cross-cultural studies by Ekman [3] and the work by Izard [17] demonstrated the universality and discreteness of subset of facial expressions, which are referred as basic or universal expressions. Generally, physical changes in face shape make the descriptor for sign based approach. The most widely-used approach is that of Ekman and colleagues, known as Facial Action Coding System (FACS) [18] (See Section 1.1.1.2 for reference).

### 1.1.1.1 Ekman's six basic emotions

The initial research conducted by Ekman determined that there are six prototypical facial expressions which are also referred as "universal" since they were found to be universal across human cultures and ethnicities [1]. These six basic facial expressions are anger, disgust, fear, happiness, sadness and surprise (see Figure 1.1), each of which expressions is performed by its exclusive or common activities of facial muscles.

- Anger - symbolised by lid and lip tightening, eyebrow lowering and eyes bulging
- Happiness - symbolised by lip corner raising, cheek raising and mouth opening partly
- Disgust - symbolised by upper lip raising and nose bridge wrinkling
- Fear - symbolised by the eyebrow pulling together, upper lids raising, middle of forehead wrinkling, and lips stretching horizontally
- Sadness - symbolised by inner brow pulling together, lip corner lowering and lids dropping
- Surprise - symbolised by eyebrows (both inner and outer) raising, jaw dropping slightly, mouth opening wide and eyes opening with more white exposed

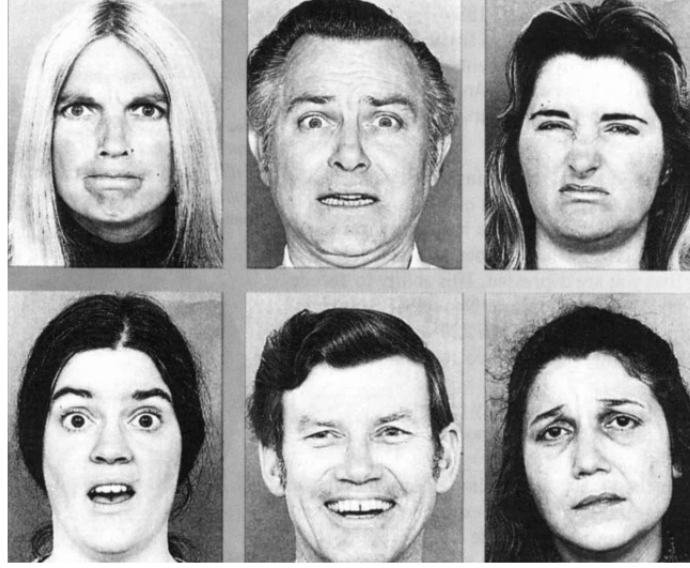


Figure 1.1: Example of expression for the six basic emotions (Left-to-right from top row: anger, fear, disgust, surprise, happiness, and sadness).

The Ekman's six basic emotion system has been extensively used in computer vision community for testing the performance of facial expression recognition system. In this research work, we also use Ekman's six basic emotion to evaluate various proposed frameworks of facial expression recognition.

### 1.1.1.2 Facial action coding system (FACS)

Facial Action Coding System (FACS) [18] is one of most widely used tools for describing facial muscle action. There are 46 Action Units (AUs) (see Figure 1.2) and Action Descriptors (ADs) defined for describe the facial expressions in FACS, where AUs encode the action (contraction or relaxation) of one or more muscles, while ADs which are different from AUs are unitary movements involving the action of several muscle groups.

Figure 1.3 shows AUs in the upper face, in the lower face, and AUs that cannot be classified as belonging to either the upper or the lower face.

## 1.1. BACKGROUND KNOWLEDGE

---

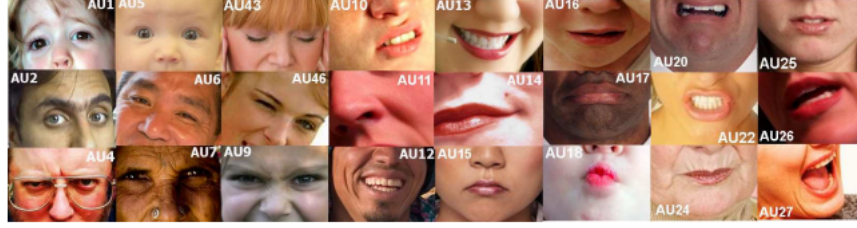


Figure 1.2: Example of AUs which defined in FACS

AU 1+2	AU 1+4	AU 4+5	AU 1+2+4	AU 1+2+5
AU 1+6	AU 6+7	AU 1+2+5+6+7	AU 23+24	AU 9+17
AU 9+25	AU 9+17+23+24	AU 10+17	AU 10+25	AU 10+15+17
AU 12+25	AU 12+26	AU 15+17	AU 17+23+24	AU 20+25

Figure 1.3: Facial Action Units.

AUs are considered to be the smallest visually discernible facial movements. They are atomic, meaning that no AU can be split into two or more smaller components. Any facial expression can be uniquely described by a combination of AUs.

### 1.1.1.3 FACS AU combinations and intensity

As AUs represent the "atoms" of facial expressions, multiple AUs often occur simultaneously. Out of 46 AUs 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face (see Figure 1.4). When AUs occur in combination, they may be additive where the combination does not change the appearance of the constituent AUs, or non-additive where the appearance of the constituents does change [19]. So far, about 7000 valid AU combinations have been identified within the FACS framework.

Further relationships among multiple AUs exist as well. For instance, in certain AU combinations, the dominant AU may completely mask the presence of another subordinate action unit. For such combinations, special rules have been added to FACS so that the subordinate AU is not scored at all [20].

### 1.1. BACKGROUND KNOWLEDGE

---










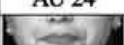


Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 1.4: Some examples of Action Units. Action Units are atomic facial muscle actions described in FACS.

In addition to determining which AUs are contained within the face, the intensity of each AU present must also be ascertained. Intensity is rated on a scale from A (least intense) through E (most intense). Criteria for each intensity level are given in the FACS Manual for each AU [18].

#### 1.1.1.4 Drawbacks of FACS

FACS codes often reveal unnecessary details that can hamper facial expression recognition approaches. The sheer number of combinations (7000 AU combinations) can lead to a bad generalisation performance as it is virtually impossible to have access to a training database that covers all possible AU combinations while featuring a sufficient number of instances of specific facial expressions [21].

Also, the problem with using FACS is the time required to coding every frame of the video. FACS was envisioned for manual coding by FACS human experts. It takes over 100 hours of training to become proficient in FACS, and it takes approximately two hours for human experts to code each minute of video [22]. Aforementioned drawbacks of FACS make it impractical for real life scenarios.



### 1.2 Objectives and motivation

Due to the limitations of FACS aforementioned, this thesis focuses on analysing and recognising six basic facial expression directly not using FACS, and addressing some challenging problems of current research in facial expression recognition. In particular, a few novel techniques [20; 23; 24] were developed to represent facial features in spatio-temporal domain, aiming at increasing the robustness and effectiveness of facial representation. Inspired by these research, we attempt to address some problems exist in spatial based facial feature representation in spatio-temporal domain (considered as 3 dimensional (3D)) not just in spatial domain.

Facial expression is one of important and effective means of non-verbal communication. When people communicate face to face, facial expression has been considered as one of useful tools in assisting a listener to infer the intention of speaker or providing feedback to the speaker even when the listener is silent [12]. To some extent, a facial expression contains rich information of personal behaviour, which is a complex reflection of human emotion, mental state and health state. Thus using computer technology to understand and analyse facial expression could change the relation between human and computer significantly, which plays an important role in realising harmonious human and computer interaction (HCI) [25].

As is widely known that it is easy for people to recognise facial expression of others when people communicate face to face. But for computer, it is a difficult task to recognise facial expressions of a person [26]. This is because expressions are displayed along with complex activities of a specific set of facial muscles, where each expression is affected by one or combination of several facial muscles [2]. It is really challenging to construct an exact mathematics model to describe the activities of these muscles. In addition, a change of facial expression normally reflects the movement of a number of facial landmarks (e.g. mouth corner, lips, eyebrows, etc.), and current computer technology is not good enough to localise the accurate locations of these landmarks, which leads to a failure in accurately detecting the activities of facial muscles [24]. Another main reason is that different people perform facial expressions in different ways, which means even the same expression might differ when performed by two different persons. Furthermore different expressions of a person do not correspond to obvious emotion [27]. For example, the opening of mouth does not always means smiling, but could be surprise or sadness. Some other factors also affect facial expression such as age, race, gender, hair, etc. Therefore, it is difficult for a computer to use a general criteria to categorise various expressions, and it is a challenging task to use computer for analysing and recognising facial expressions in the area of affective computing [26].

In recent years, facial expression recognition has been an active research topic, which has drawn more and more attention [19]. The following are some reasons. First, the development of technology on relevant research areas encourages the advancement of facial expression recognition. In the past decades, face recognition [28; 29], face detection [30; 31], face tracking [32; 33] have made a huge progress, e.g., the accuracy and robustness of face detection and tracking methods have been significantly increased. Second, the wide range applications of facial expression recognition in terms of both academic and business has accelerated the development of facial expression recognition [34; 35]. For example, a computer game is more attractive and realistic if the game is capable of reacting in real-time according to the expressions of gamer (e.g., anger and happiness). In some surveillance systems of different environments such as cars, planes, factories, etc., certain passive mental states (e.g., fatigue) can be detected immediately by a facial expression recognition system so as to give timely alarms and thus avoid the occurrence of an accident [36]. In the field of linguistics, facial expression can be combined with lip reading to help people who have problems with hearing to communicate with others. In addition, facial expression analysis has extensive applications on security environment and medical treatments [35]. Thus, due to their potential applications and promising future, facial expression recognition can become a useful clue to solve problems of other relevant research fields, and also should be paid more attention.

### 1.3 Contributions

This thesis proposes three novel facial expression recognition frameworks: Framework 1 (presented in Chapter 3), Framework 2 (presented in Chapter 4) and Framework 3 (presented in Chapter 5).

Framework 1 uses sparse representation to extract features from patches of a face to improve the recognition performance, where part-based methods which are robust to image alignment are applied. In addition, the use of sparse representation reduces the dimensionality of features, and improves the semantic meaning and represents a face image more efficiently.

Since a facial expression involves a dynamic process, and the process contains information that describes a facial expression more effectively, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Thus, Framework 2 uses two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on PHOG (pyramid histogram of gradient) to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. The

framework views an image sequence as a spatio-temporal volume, and uses temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. Specifically, spatial based descriptor representing spatial local shape is extended to spatio-temporal domain to capture the changes in local shape of facial sub-regions in the temporal dimension to give 3D facial component sub-regions of forehead, mouth, eyebrow and nose. The descriptor of optical flow is also employed to extract the information of temporal. The fusion of these two descriptors enhance the dynamic information and achieves better performance than the individual descriptors.

Framework 3 also focuses on analysing the dynamics of facial expression sequences to represent spatial-temporal dynamic information (i.e., velocity). Two types of features are generated: a spatio-temporal shape representation to enhance the local spatial and dynamic information, and a dynamic appearance representation. In addition, an entropy-based method is introduced to provide spatial relationship of different parts of a face by computing the entropy value of different sub-regions of a face.

## 1.4 Structure of thesis

The thesis begins with an overview of facial expression recognition in Chapter 2, which explains the existing techniques, challenges and problem space. Special focus is made on reviewing existing methods for facial expressions recognition. The datasets which are widely used in the field of facial expression recognition are also introduced in this chapter.

Chapter 3 provides a detailed explanation of Framework 1 which is based on patch methods. A novel method of detecting face regions is also introduced in this chapter.

An extension to 2D spatial descriptor that uses two types of dynamic information is presented in Chapter 4. An analysis on the contribution of different facial subregions using the Framework 2 is also presented in this chapter.

Framework 3 is presented in Chapter 5, which exploits a spatio-temporal feature based on local Zernike moment in the spatial domain using motion change frequency. In addition, the design process of a dynamic feature comprising motion history image and entropy is introduced in this chapter.

Chapter 6 provides the concluding remarks of the thesis and outlines the current limitations together with future research directions.

## Chapter 2

# Literature review on Facial expression recognition

This chapter provides literature review on facial expression recognition, which introduces techniques in the area of facial expression recognition, the widely used datasets, current challenges and problem space. Section 2.1 briefly covers core components of a system that recognises facial expression. Section 2.2 describes characteristics of different datasets used for evaluating and benchmarking different facial expression recognition algorithms. Section 2.3 and Section 2.4 provide multi-modal expression analysis and the problem space for facial expression recognition, respectively.

### 2.1 Techniques in facial expression recognition systems

In general, a facial expression recognition system comprises three modules: face acquisition, feature representation and classification.

The stage of face acquisition includes face detection and face alignment, where face detection is used to find the face region in the input images or image sequences, and face alignment is applied to reduce the effect of variation in head pose and scene illumination to give a better recognition performance.

After the face is located, the next stage is to extract and represent the facial changes caused by facial expressions, which is very crucial for further processing. The recognition performance of the whole framework greatly depends on the representation extracted from face images. If the computed representation is accurate enough, using a relative simple classifier could achieve an acceptable recognition performance, which means complicated classifiers are not necessarily. How to obtain robust and effective facial features is the main focus of our work. Thus, attention is

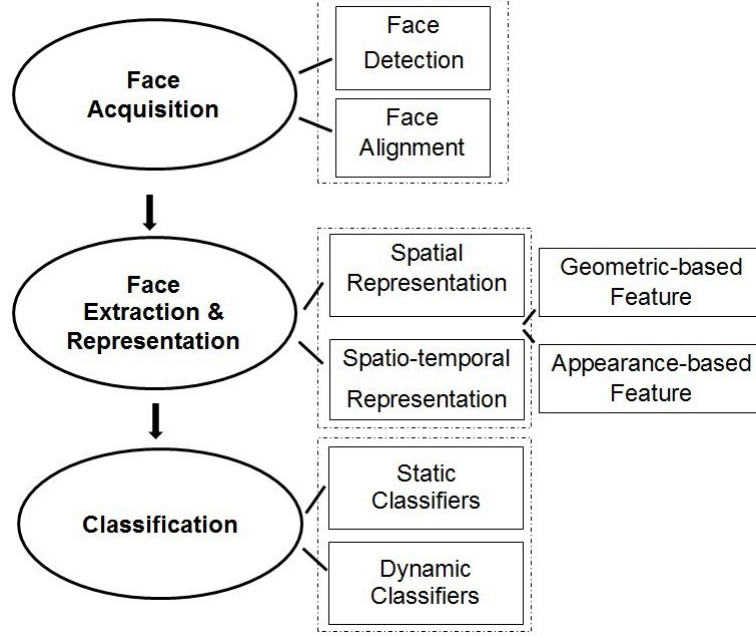


Figure 2.1: Basic structure of facial expression recognition system

paid to extracting facial features to improve the recognition performance (in terms of recognition rate) of facial expression recognition.

The classification of a detected facial expression is the last stage of facial expression recognition systems. The facial changes can be identified as facial action units or prototypical emotional expressions. Our work focuses on the task of recognising prototypical emotional expressions.

In the following subsections (Section 2.1.1: face acquisition which includes face detection and alignment, Section 2.1.2: feature extraction and Section 2.1.3: expression classification), the literature on all of the three modules will be briefly reviewed.

### 2.1.1 Face acquisition

The first step in facial expression recognition is face acquisition which includes face detection and face alignment. Locating the face region in a image or single frame of a video sequence is referred as face detection, while locating the face region over the whole video sequence is termed as face tracking. Face detection plays an important role in facial expression recognition. Face is a kind of natural structure which has a fairly complex details variation. The challenges of detecting such natural structure are as follows[37]: (1) Due to the difference of facial appearance, expression, skin colour, different faces have different patterns; (2) a basic face is usually occluded by

hair, glasses or beard; (3) as a 3D object, face image is inevitably affected by the shadow of light produced. Therefore, addressing these problems above can provide a clue to generate an effective face detection framework which detects face more accurately and quickly. Methods for face detection can be grouped into four categories [38]: knowledge-based methods, feature-based methods, template matching methods, and appearance-based methods. Face alignment based on whole face is usually performed by detecting facial landmark and using their location to compute a global transformation (e.g. Euclidean, affine) that maps an input face to a basic face. The transformation can be computed using Active Appearance Models (AAM).

### 2.1.1.1 Face detection

Face detection methods can be classified into four categories [38]: knowledge-based methods, feature-based methods, template matching methods, and appearance-based methods.

**Knowledge-based methods:** knowledge-based methods use pre-defined rules or models to determine a face based on general human knowledge. Usually, the rules capture the relationships between facial features. It is trivial to find simple rules to describe the features of a face and their relationships. For example, face appears in an image with two eyes (usually), a nose and a mouth. To describe relationship between features, distance and relative position are good metric.

In [30], Yang and Huang proposed a face detection method using hierarchical knowledge, which they proposed three level system, where at first level all possible face candidates are found. The rules at a higher level are general descriptions of what a face looks like while the rules at lower levels rely on details of facial features. Motivated by the simplicity of the approach proposed by Yang and Huang [30], Kotropoulos and Pitas [39] proposed face detection algorithm which extends their method. Kotropoulos's method [39] makes computationally complex algorithm of Yang's [30] much simpler. Mekami and Benabderrahmane [40] proposed algorithm that not only detects face but also its inclination. They used Adaboost learner for face detection. For the calculation inclination, they used an eyes detector. Then the line passing through the two eyes is identified, and the angle to horizon is calculated.

Problem with this approach is the difficulty in translating human knowledge into rules. If the rules are detailed (i.e., strict), they may fail to detect faces that do not pass all the rules. If the rules are too general, they may give many false positives. Moreover, it is difficult to extend this approach to detect faces in different poses since it is challenging to enumerate all the possible cases. On the other hand, heuristics about faces work well in detecting frontal faces in uncluttered scenes [30].

**Feature-based methods:** Feature-based methods aim to find some facial features that still exist even when pose or light conditions changes, and use them to obtain the exact location of face. These facial features include skin colour and texture.

In the colour images, the skin colour is an important feature of human face, and using skin colour as a cue feature is an effective way to detect human face. In the recent years, researchers have attempted to detect skin colour from colour image in moderate light [41; 42]. Lighting compensation methods and non-linear colour transformation has been used for detecting skin regions, and maps of eyes, mouth, and nose can be generated to verify face candidate [43]. In some cases where the pixels between skin and non-skin overlap in many colour space, textural and spatial features are employed for modelling skin [44]. Tan proposed a skin detection algorithm along with eye detection combining Gaussian model and 2D histogram [45] for detecting human skin automatically in colour image. However, all of these existing methods focus on generating skin model and training parameters, which increase detection performance at the cost of calculations. Also, the performance of detection is affected by several challenging factors: illumination variation, different ethnic groups and complex background. The variation in light source or in the illumination level (indoor, outdoor, shadows, etc.) might cause the change in the colour of the skin. Normally, because of the non-plane shape of the facial features, the dark shadow in the face may result of directional lighting that blackened some facial parts. Sometimes, the reflection of strong lighting may lead to a bright spot in a face. People from different ethnic groups (race) have various skin colour appearance, and they also have different face structures. For instance, Asian people have yellow skin colour, while Europeans have white skin colour. Complex background is another challenge. In the real world, furniture, clothes, hair, etc. may bring skin-like colour, which causes the skin detector to make false detection.

Some distinct texture in human face can be used for separating them from different similar subjects. In [46], Augusteijn and Skufca proposed a method inferring the presence of a human face by using the identification of face like texture (skin, hair and others) which are computed using second order statistical features (SGLD) on sub regions of  $16 \times 16$ . A cascade correlation neural network [47] was used for classifying texture, and a Kohonen self-organising feature map [48] was exploited to generate clusters for different texture classes.

**Template matching based methods:** template matching based methods manually predefine a general face pattern (model) using a function. The parameters change according to different input images, which is com-

puted for eyes, nose, mouth and face contour independently. Template matching based methods is easy to implement with low computational complexity. However, they are not good when dealing with changes in scale, pose and shape and subsequently deformable templates have been presented to obtain the invariance of shape and scale. Sinha proposed a Ratio Template Algorithm for the cognitive robotics project at MIT [49]. This algorithm was used to develop a system that located the eyes by first detecting the face in real environment [50]. In [51], Anderson and McOwen applied "Golden Ratio" to the Ratio Template, and designed a modified version referred as the Spatial Ratio Template tracker, which showed a better performance under different illumination. However, the limitations of predefined templates based methods are their inadequacy to deal with variation in scale, pose, and shape.

**Appearance-based methods:** unlike template-matching methods which rely on a predefined template or model, appearance-based methods use large numbers of examples (images of faces and or facial features) depicting different variations (face shape, skin colour, eye colour, open closed mouth, etc.) Face detection in this case can be viewed as a pattern recognition problem with two classes: face and non-face [40]. In general, appearance-based methods rely on techniques from statistical analysis and machine learning to find the relevant characteristics of face and non-face images [38]. Seminal work in appearance-based methods for face detection are based on eigenfaces [52], neural networks [53], support vector machines [54] and hidden Markov models [31].

Adaboost (proposed by Freund and Schapire [55]) has also been used by several researchers to create robust system for detection of objects in real time [38]. Papageorgiou et al. used this algorithm to detect pedestrians using the Haar wavelet to extract discriminating features in [56]. Inspired by the algorithm proposed by Papageorgiou et al., Paul Viola and Michael Jones [57] proposed an algorithm for face detection.

In our research work, we used Viola-Jones object detection algorithm for detecting face and salient facial regions as it is the most cited and considered the fastest and most accurate pattern recognition method for face detection [38]. Viola-Jones object detection algorithm is explained below.

Viola-Jones detection method incorporates the following four key concept:

1. *Haar-like features:* The first contribution of Viola and Jones algorithm is computational simplicity for feature extraction. The features used by their method are called Haar-like features.



2. *Integral image*: To rapidly compute Haar-like features Viola and Jones proposed intermediate representation for the image, called integral image. The integral value for each pixel location  $(x, y)$  is the sum of all the pixels above it and to the left of  $(x, y)$  inclusive (refer Equation 2.1):

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'), \quad (2.1)$$

where  $ii(x, y)$  is the integral image and  $i(x, y)$  is the original image.

Using the integral image, any rectangular sum can be computed in four array references (see Figure 2.2). For example, the value of an integral image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is  $A + B$ , and location 4 is  $A + B + C + D$ . The sum within D can be computed as  $4 + 1(2 + 3)$ . Clearly, the difference between two rectangular sums can be computed in eight references. Since the two-rectangle features defined above involve adjacent rectangular sums they can be computed in six array references, eight in the case of the three-rectangle features, and nine for four-rectangle features [57].



Figure 2.2: The sum of the pixels within rectangle D can be computed with four array references. The value of the integral image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is  $A + B$ , At location 3 is  $A + C$ , and at location 4 is  $A + B + C + D$ . The sum within D can be computed as  $4 + 1 - (2 + 3)$ .

3. *AdaBoost method*: Viola and Jones selected an AdaBoost training algorithm proposed by Freund and Cshapire [55]. This algorithm is used to determine the presence Haar-like feature by setting threshold levels. Threshold value is used to determine the presence of a Haar-like feature (the sum of pixels values in the shaded rectangle is subtracted from the white rectangle). If the difference is above a threshold, that feature is regarded to be present.

Secondly, within any image sub-window the total number of Haar-like features is very large, far larger than the number of pixels. In order to ensure fast classification, the learning process must exclude a large majority of the available

features, and focus on a small set of critical features. To achieve this goal, the weak learning algorithm is proposed by Viola Jones which selects the single rectangle feature which best separates the positive and negative examples. For each feature, the weak learner determines the optimal threshold classification function, such that the minimum number of examples are misclassified [57].

Learning algorithm combines many weak classifiers to create one strong classifier, where weak means the classifier only gets the right answer a little more often than random guessing would. Then, this learning algorithm selects a set of weak classifiers to combine and assigns a weight to each. This weighted combination is the strong classifier.

4. *Cascades of classifiers*: Viola and Jones introduced a method for combining successively more complex classifiers in a cascade structure for increased detection performance while radically reducing computation time. The rationale behind this is to construct boosted classifiers which reject many of the negative sub-windows while detecting almost all positive instances. Simpler classifiers are used to reject the majority of sub-windows before more complex classifiers are called upon to achieve low false positive rates [57]. Each stage is only required to eliminate slightly more than 50% of false detection as long as it kept the positive hit rate close to 100%.

### 2.1.1.2 Face alignment

Face alignment is a fundamental step for facial expression recognition, which is designed to reduce the affect resulting from the variation of head pose and face geometry. The performance of face alignment have a direct impact on further process and the performance of recognition. Basically, the alignment can be divided into whole face, part and point alignment based on the output of process of alignment.

Similar to whole face alignment, part based alignment use AAM technique, where the facial parts are localised as fixed size patches around detected landmarks. In some case, faces might be warped into a reference frontal face model non-rigid alignment before patches are cropped (e.g. [58; 59]). Techniques that perform part detection to localise each patch individually can also be used [60].

Point based alignment is useful for shape descriptor, which involves the localisation of fiducial (stable) points. AAM is also used for point alignment. As the accuracy of locations of facial points is important for shape representations, it is necessary to validate the feature detectors across expression variation [61]. Due to the demand for real time application, points in a video sequence can be also aligned

by detecting points use a point detector on the first frame and then tracking them using tracking methods (e.g., Gabor based point detector [62] and particle filter [63]).

### 2.1.2 Feature representation

After aligning the detected face image, feature representation and extraction is applied to analyse and recognise facial expressions automatically. The idea behind the feature presentation aims to extract optimal feature that should minimise intra-class variation of facial expression, while maximising inter-class, while maximising inter-class variation [64]. In some case, the best classifiers might fail to gain satisfied recognition if inadequate features are exploited.

#### 2.1.2.1 Spatial representation

Spatial features encode an image or single frame from video sequences which can be referred as 2D [65]. Various spatial methods for facial expression extraction have been proposed, which can be categorised as geometric-based and appearance-based methods. Geometric-based features are represented using the information of shape and locations of facial components (including mouth, eyes, brows, nose, etc.), while the appearance features capture the information of texture of face, such as wrinkles and furrows.

**Geometric-based methods:** in the geometric-based approaches, representation using a set of facial points is one of the most frequently used, where the  $x$  and  $y$  coordinates of facial points are simply concatenated to form a feature vector as input of classification [66; 67]. There are some other methods to represent shape which are less common. Huang, et al. use the relative distances between facial landmarks instead of row coordinates to describe the shape in [68]. In [19], Tian, et al. compute the distances and angles of landmarks to represent the activities of various facial parts i.e. movements of eyebrow, opening/closing of the mouth and the state of the cheek. One of limitations of appearance-based approach is that the representation based on either the raw location of landmarks or the relative location deeply depends on the performance of points detection and alignment, which means it is easily affected by the error of previous step (i.e. face detection and alignment). However, geometric-based approach is relatively robust to the variations of illumination as the intensity of the pixels is not used.

As mentioned earlier, geometric features present the shape and locations of facial components (including mouth, eyes, brows, nose). Thus, the motivation for employing a geometry-based method is that facial expressions affect the relative

position and size of various facial features, and that, by measuring the movement of certain facial points, the underlying facial expression can be determined. In order for geometric methods to be effective, the locations of these fiducial points must be determined precisely; in real-time systems, they must also be found quickly. The exact type of feature vector that is extracted in a geometry-based facial expression recognition systems depends on a) which points on the face are to be tracked, b) whether 2D or 3D locations are used, c) the method of converting a set of feature positions into final feature vector.

The typical examples of geometric based methods for facial expression recognition are those of Pantic and her colleagues [69], who used a set of facial characteristic points around the mouth, eyes, eyebrows, nose, and chin.

Research has been done with success in recent times to combine features extracted using appearance-based methods and geometric feature-based methods [58; 70]. The key problem with geometric methods is to precisely locate the landmark and track it. In the real applications, due to the pose and illumination variations, small resolution input images, and the noise from the background, it is still very hard to precisely locate the landmarks.

**Appearance-based methods:** existing appearance-based approaches for facial expression recognition encode low-level or high-level information, where the former one which is more popular interprets expressions using low-level histograms and Gabor descriptor, while the latter one is encoded using sparse coding [71; 72; 73] or Non-Negative Matrix Factorisation (NMF) [74; 75]. Some part-based methods [76; 59] attempt to extract appearance information by splitting face region into a number of local face components. Brief overview of the above mentioned representations are given as follows.

Gabor representation is an representative feature vector of such approach that describe the local appearance models of facial expressions, which is applied in different recognition frameworks (i.e. the winner of FERA AU detection [77; 78] and AVEC [79]). The Gabor decomposition of an image is computed by convolving the input image with a set of Gabor filters, which can be tuned to a particular frequency  $k_0 = (u, v)$ , where  $k = \|k_0\|$  is the scalar frequency and  $\varphi = \arctan(\frac{u}{v})$  is the orientation [80; 81]. Gabor filters accentuate the frequency and orientation respectively. A Gabor filter can be represented in the space domain using complex exponential notation as:

$$F_{k_0} = \frac{k_0^2}{\sigma^2} \exp(-\frac{k_0^2 x^2}{2\sigma^2}) (\exp(ik_0) - \exp(-\frac{\sigma^2}{2})), \quad (2.2)$$

where  $x = (x, y)$  is the image location and  $k_0$  is the peak response frequency. An

example of a Gabor filter is given in Figure 2.3, which shows the absolute value (left), real component (middle) and imaginary component (right) of the filter in the space domain.

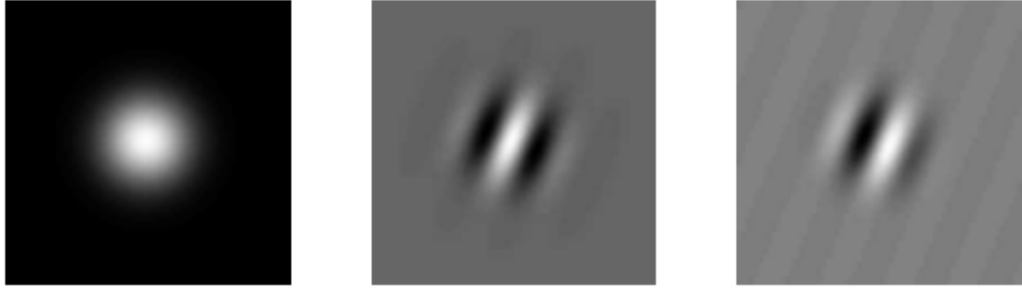


Figure 2.3: The left, middle, and right graphics above show the absolute value, and the real and imaginary components of a sample Gabor filter.

For expression analysis, a filter bank of multiple Gabor filters tuned to different characteristic frequencies and orientations is used for feature extraction. The combined response is called a jet. Filter banks typically span at least 6 different orientations and have frequencies spaced at half-octaves. Prior to classification, the extracted features are usually converted into real numbers by calculating the magnitude of the complex filter response.

Due to the good performance on describing facial texture, Gabor feature are utilised for recognising facial expression. Lyons et al. [82] developed a recognition framework by computing Gabor responses over facial mesh points, which achieved recognition rate of over 90% by using linear discriminant analysis and classical nearest neighbour classifier. In the work of [83; 84], Tian et al. also applied Gabor features for recognising FACS AUs, where they extracted Gabor bank along with a number of landmarks (i.e. the corner of eyes, forehead, the corner of eyebrows, etc.). The developed system gained satisfying recognition rate for recognising AUs by using three layer neural network. Donato et al. [85] proposed a recognition framework based on whole face region using Gabor filters. They encoded face image using a filter bank of 5 frequencies and 8 spatial orientations, which achieved a good recognition rate of over 95% on several AUs. Littlewort et al. has shown a high recognition accuracy (93.3% for Cohn-Kanade facial expression database) using Gabor features. They proposed to extract Gabor features from the whole face and then selected the subset of those features using AdaBoost method. In [86], Littlewort et al. computed Gabor features on difference image instead of raw image and use support vector machines to classify AUs 6 and 12 corresponding to smile expression. However, Gabor features still have one main limitation that they produce

a large dimensionality of feature which leads to a high computational cost in terms of time and memory, especially if they are applied to a wide range of frequencies, scales and orientations of the image features.

In recent years, some frameworks based on low-level histogram are proposed for facial expression recognition, which achieves satisfied performance [87; 88; 89; 90; 91]. Low-level features are usually extracted from local regions, and encode the local texture information by computing histogram. Then, the histograms on each local region are concatenated into feature vector by different normalising and pooling strategies. The final representation are the concatenation of all local histogram. Low-level representation has tolerance to variations of illumination to she extent, since they encode the local regions. Also, the use of normalisation is able to deal with the global variation of illumination, and pooling can reduce the affect of alignment errors. However, low-level representations take no account of semantic information and con figural information. Several representative methods are briefly introduced below [90].

Local Binary Pattern (LBP) were initially proposed for texture analysis [92], but recently they have been successfully used for face and facial expression analysis [87; 88]. The most important property of LBP features are their tolerance against illumination changes and their computational simplicity. The operator labels the pixels of an image by thresholding the  $3 \times 3$  neighbourhood of each pixel with the centre value and considering the result as a binary number. Then the histogram of the labels can be used as a texture descriptor. Formally, LBP operator takes the form:

$$LBP(x_c, y_c) = \sum_{n=0}^7 s(i_n - i_c) 2^n, \quad (2.3)$$

The operator later was extended to use neighbourhood of different size [93]. Using circular neighbourhoods and bilinear interpolating, the pixel values allow any radius and number of pixels in the neighbourhood. See 2.4 for examples of the extended LBP operator, where the notation  $(P, R)$  denotes a neighbourhood of  $P$  equally spaced sampling points on a circle of radius of  $R$  that form a circularly symmetric neighbour set.

Shan [89] et al. applied the LBP features on facial expression recognition and also got promising performance. Zhao et al. [88] proposed to model texture using volume local binary patterns (VLBP) an extension to LBP, for expression recognition. Average FER accuracy of 96.26% was achieved for six universal expression with their proposed model on CK [94] facial expression database. Due to

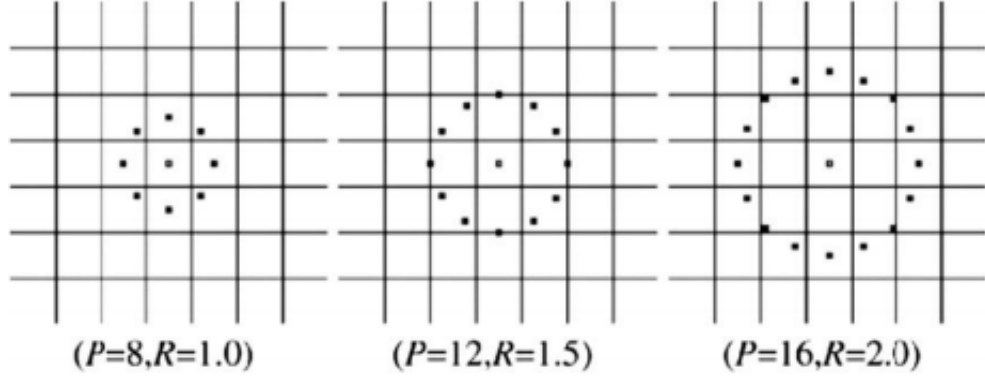


Figure 2.4: Three examples of the extended: the circular (8; 1) neighbourhood, the circular (12; 1) neighbourhood, and the circular (16; 2) neighbourhood, respectively

both spacial and temporal information is considered in VLBP, it got better result comparing with the traditional LBP.

Histogram of gradient (HOG) [95] was originally developed for person detection and object recognition. In [96], HOG descriptor are extracted from face image using a dense grid, and are used for face recognition. Histograms of Oriented Gradients are generally used in computer vision, pattern recognition and image processing to detect and recognise visual objects (i.e. faces). We propose to use HOG descriptors because we need a robust feature set to discriminate and find faces under difficult illumination backgrounds, wide range of poses, etc., by using feature sets that overcome the existing ones for face detection.

The ideas behind HOG are edge orientation histogram, SIFT descriptor [97] and shape context. They are computed on a dense grid of cells that overlap local contrast histogram normalisations of image gradient orientations to improve the detector performance [95]. So that, this feature set performs very well for other shape based object classes (i.e. face detection) because of the distribution of local intensity gradients, even not processing any knowledge of the corresponding gradient [97].

To compensate the illumination, histogram counts are normalised by accumulating a measure of local histogram energy over the connected regions, then use the results obtained to normalize all cells in the block (e.g., size 2) and finally, the combination of these histograms represents the HOG descriptor, which can yield a better robustness to variation in illumination. The HOG descriptor is affected by several of its parameters: the choice of gradient operator; the binning of orientation; the normalisation scheme; and the geometrical shape of blocks for subdivision

(i.e., square and rectangular). Dalal and Triggs gave a detailed and systematically implementation study on the effect of the various choices on descriptor performance in [98].

Pyramid HOG (PHOG) was proposed in [99], which is a spatial shape descriptor and got its inspiration from the works of Dalal et al. [100] on histograms of oriented gradients and Lazebnik et al. [101] on spatial pyramid matching. It represents an image by its local shape and the spatial layout of the shape. The idea of PHOG is shown in 2.5. Process for the PHOG feature extraction are explained as follows:

- 1) Canny edge operator is applied to extract contours from the given stimuli.
- 2) Then, the image is divided into spatial grids by iteratively doubling the number of divisions in each dimension.
- 3) Afterwards, a histogram of orientation gradients are calculated on each edge point using  $3 \times 3$  Sobel mask without Gaussian smoothing and the contribution of each edge is weighted according to its magnitude. Within each cell, histogram is quantized into  $N$  bins. Each bin represents the accumulation of number of edge orientations within a certain angular range.
- 4) To obtain the final PHOG descriptor, histograms of gradients at the same levels are concatenated. The final PHOG descriptor is a concatenation of HOG at different pyramid levels. Generally, the dimensionality of the PHOG descriptor can be calculated by:  $N \sum_l 4^l$ .

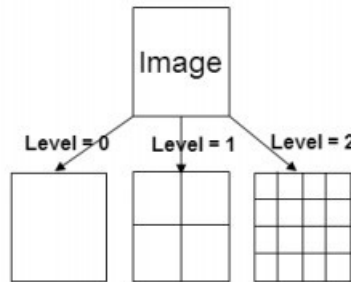


Figure 2.5: PHOG descriptor of an image.

In recent years, local moment based methods has been applied for facial expression recognition, which achieves a good performance [102]. Classical image moments can be categorised into geometric moments, complex moments and orthogonal moments. Although easy to use, the large values of geometric moments



are their main limitations leading to numerical instabilities and sensitivity to noise. Complex moments are defined similarly to geometric and have been used to describe the shape of a probability density function and to measure the mass distribution of a body. Hu moments exhibits translation, rotation and scaling invariance, and has been applied in many areas [103]. Orthogonal moments are projections of a function onto a polynomial basis. ZMs employ complex Zernike polynomials as its moment basis set [104], and have been used to recognise facial expressions [105]. The rotation invariance of Zernike-based facial features is discussed in [106; 29]. QLZM is used in [102] for recognising facial expressions. However, ZM has its shortcomings, namely it is a low level histogram representation which ignores the spatial relations (i.e., configure information) among the different facial parts. Also, ZMs only describe the texture information in each frame of image sequences, and do not capture any dynamic information.

All aforementioned representation describe local texture, and their features encode the distribution of edges. Recent methods aim instead to obtain higher-level representations to encode features that are semantically interpretable. Two such representations are Non-Negative Matrix Factorisation (NMF) and sparse coding.

NMF methods decompose a matrix into two non-negative matrices. The decomposition is not unique and can be designed to have various semantic interpretations. One NMF-based technique is Graph-Preserving NMF (GP-NMF) [75], which decomposes face into spatially independent components through a spatial sparseness constraint [107]. Another NMF-based method is Subclass Discriminant NMF (SD-NMF) [74], which represents an expression with a multimodal projection. Unlike GP-NMF, SD-NMF does not explicitly enforce decomposition into spatially independent components. The basis image provided [74] suggest that the information encode can be holistic componential or configural. NMF methods creat a number of basis images, and the features of NBF representations are the coefficients of each basis image. The method performs minimisation to compute the coefficients, therefore its computational complexity varies based on the optimisation algorithm and the number and size of basis images.

Sparse coding based methods which exploit sparsity in pattern classification have been proposed recently, which has shown promising performance. Wright et al. [108] used sparse representation for face recognition to find a face representation using a sparse linear combination of an over-complete dictionary. More specifically, the sparse-representation based classifier (SRC) in [108] is a novel algorithm for robust face recognition, which can address the problems of face occlusion, corruption and disguise. The idea behind SRC is representing the test face image using a

small number of atoms which is chosen from an over-complete dictionary which consists of all training samples. The sparsity constraint of the coding coefficients is employed to ensure only a few samples from the same class of the query face have distinct non-zero values, whereas the coefficients of other samples are equal or close to zero [108]. The sparsity of the coding coefficient can be directly measured by  $l_0$ . It is shown that if there are sufficient training samples for each facial class, it is possible to represent a test image sample as a linear combination of only those training samples that belong to the same facial class [108]. The sparse representation they proposed achieved high responses to the face image of the same identity class and low responses to the other classes. Motivated by the work of Wright et al., Zafeiriou and Petrou [73] used sparse representation to recognise facial expressions. In their paper, it is shown that the performance of sparse representation using the original images directly for facial expression recognition is not satisfactory, and the sparse coefficient vectors are much more related to the facial identity than to facial expression. They argue that to achieve more efficient representations, person-independent representation should be applied such as difference image and facial grids. However, their methods still use full face image which might not be able to handle occlusion. SRC proposed by Wright et al. [108] was applied for the purpose of facial expression recognition in [109; 110; 111]. Most of existing methods using sparse representation made use of features which describe intensity changes of face appearance (i.e., appearance based feature). In [109], Huang et al. combined the SRC with LBP features and shape of image [112], which showed the SRC using LBP achieved better recognition performance than SRC using image raw pixels. [110] showed that SRC performed better than using nearest neighbour [113] and support vector machine classifiers [112], independent of features used (e.g., raw pixel, LBP, Gabor wavelet, etc.) However, In comparison with extracting facial features from the whole image, patch based methods which extract facial features by dividing a face image into several sub-regions (patches) have shown promising performance as reported in [114; 115; 116; 117].

Most existing methods exploit the whole face image, which leads to some limitations on face alignment, and they are not able to handle the problems associated with occlusions. In the case of using the whole image, small misalignment might cause displacement of the sub-region locations which increase the classification error. The size and shape of facial organs from different persons are not the same. Also, the same facial position are not always present in one particular block in all images. In comparison with extracting facial features from the whole image, patch based methods which extract facial features by dividing a face image

into several sub-regions (patches) have shown promising performance as reported in [118; 119; 36; 60]. In [118], face image is divided into several  $7 \times 6$  sub-regions and  $7 \times 6 \times 59$  dimensional local features are extracted. The discriminative LBP histogram bins are then selected by using Adaboost technique for optimum classification. In [119], the face image is divided into 64 sub-regions, and the common facial patches which are active for most expressions and special facial patches which are active for specific expressions are explored. Using a multi-task sparse learning method, features of a few number of facial patches are used to classify facial expressions. Song et al. [36] used eight facial patches based on specific landmarks positions to observe the skin deformations caused by expressions. The authors used binary classifiers to generate a Boolean variable for presence or absence of skin wrinkles. However, these patches do not include the texture of lib corners, which is important for expression recognition. In [60], the authors extracted Gabor features of different scales from the face image and trained using Adaboost to select the salient patches for each expression. However, the size and position of a salient patch is different when trained with different databases. Therefore, a unique criteria cannot be established for recognition of facial expressions in unseen images. Several similar part based methods have been proposed in [120; 121; 122].

**Combination of geometric-based and appearance-based methods:** recent trend on feature extraction has attempted to combine geometric-based feature and appearance-based feature. In [123], Kotsia et al. proposed a method which fused texture information and shape information based on deformed Candide facial grids corresponding to the facial expression, and use SVM for recognising six basic facial expression and AUs. The recognition rate on the CK+ dataset is over 92% when recognising the six basic expression and neutral. Dhall et al. [90] designed a system which encoded shape information using PHOG and appearance information using LBP for automatic emotion recognition competition FERA 2011. The experiment results conducted on the SSPNET GEMEP-FERA datasets [124] performed better than the baseline results. Using shape features in conjunction with appearance features has proved useful and promising as the combination of two different types of features is cable of employing local, holistic and configural information which is in accordance with the behaviour of human vision system [65].

### 2.1.2.2 Spatio-temporal feature representation and extraction

Spatio-temporal descriptors treat a range of frames with a temporal windows as a single entity, and capture the temporal variation of appearance and geometric representations. Since various expressions are displayed dynamically, and dynamics

are one of important factors for distinguishing between deliberated and spontaneous expressions, spatio-temporal representation which is able to capture the temporal variation might be a useful way to analyse the variation of inter-class expressions. Similar to their spatial counterpart, spatio-temporal representation also is divided into appearance-based and geometric-based.

Three Orthogonal Plane (TOP) is a popular strategy for extending 2-dimensional (2D) spatial representation into 3-dimensional (3D) spatio-temporal one. More specifically, TOP based features are obtained over three orthogonal plane: the spatial XY plane, the horizontal temporal XT plane and the vertical temporal YT plane. A video sequence can be consider as a stack of XY slices (image) in the temporal dimension, and similarly for XT and YT slices but in the Y and X dimension, respectively, and the low-level features are extracted over each slices. The TOP strategy is originally proposed in [88], which used for extending LBP to LBP\\_TOP. Several LBP\\_TOP descriptors are then used for recognising basic expression [125] and facial AUs [126; 127]. Inspired by the work of [88], LPQ\\_TOP descriptor is proposed by extending LPQ in spatial domain to spatio-temporal domain, which is also used for AU recognition. LGBP\\_TOP is another TOP based method, which is the spatio-temporal version of local gabor binary pattern (LGBP). As shown in the relevant papers, all of these TOP based spatio-temporal methods outperform their spatial counterparts. This is because these spatio-temporal descriptors not only inherit their merits (i.e. robustness to variation of illumination and componential information), but exploit the temporal variation.

An alternative way to extract spatio-temporal appearance representation is using the convolution of spatio-temporal filters. Two typical methods are Independent Component (IC) filter [128] and Gabor filter [129]. For Gabor representation, Wu et al. [129] explored Gabor motion energy filters (GME) which is based on spatial Gabor energy filter (GE) for dynamic facial expression. Experiments in [129] show GME achieve better performance than its spatial counterpart, especially on those expressions with low intensity. In [128], Long et al. use independent component analysis to learn spatio-temporal filters, and construct representation based on learned filters.

Spatio-temporal geometric representation tracks a set of fiducial points over image sequence, and use the movement of these points to describe the shape variation of facial regions [88; 130]. In general, the raw coordinates, the magnitude and angle between pairwise points are obtained, whose differences between neutral and other expressions are the input feature of classification. Some features such as the distance between upper and button lip (shape of mouth), the shape of eye and

the distance between eyebrow and nose describe componential information. One of limitations of spatio-temporal geometric representation is that they require accurate location of tracked points, which is sensitive to registration errors.

Optical flow algorithms are popular on action recognition, which are also used for some of early expression recognition systems [131]. Optical flow based methods track the motion of objects or several points over an video sequence, and forming a feature vector indicating the estimated magnitude and direction of motion of objects. The formed vector captures the movement (motion) of fiducial points (e.g. facial landmarks for facial expression).

Koelstra et al. [132] analysed two representation: Motion History Images (MHI) [133] and Free-form Deformation (FFD) that original for registration [134], and used them to extract features by computing the spatial and temporal displacement of pixels for the recognition of facial AUs. MHI decomposes motion-based recognition by first describing where there is motion (i.e. the spatial pattern) and then describing how the object is moving [135]. In MHI, the information of motion of an image sequence is encoded by a single image where the intensity of each pixel denotes the recent movement, and the movement between two consecutive frames of video sequence are represented by thresholded different images. Unlike common techniques of feature extraction which extracting feature from uniform facial parts, Free-form Deformation extract features from several non-uniform parts using quad-tree decomposition strategy. The advantage of using this strategy is that the facial parts with facial activity of high intensity would be assigned a larger number of smaller subregions, which enhances the local information during the process of partitioning. A temporal parameters  $\theta$  which controls the length of maximum history and the speed of movement are exploited in both methods. They do not use the computed optical flow feature directly. Instead, the histogram representation as well as vector-geometric properties (e.g. curl and divergence) are introduced to obtain feature representation. Both MHI and FFD employ componential information and temporal dynamics, and the recognition performance of the latter outperforms the former one [132]. One of the advantages of MHI is that information in a range of times may be encoded in a single frame, and in this way, the MHI spans the time scale of the human motion. In MHI, the intensity value of each image pixel denotes the recent movement, ignoring the speed of the movement. However, speed can be used to distinguish the movement of some facial parts (e.g., opening of mouth and raising of eyebrows) and the movements caused by changes of in-plane head pose or relatively stable facial parts (e.g., cheek, nose, forehead, etc.) during expressions. Entropy-based methods extract intensity information of image pixels, and have been

applied for face recognition. For example, Cament et al. [136] combined entropy-like weighted Gabor features with the local normalisation of Gabor features. Chai et al. [137] introduced the entropy of a facial region, where a low entropy value means the probabilities of different intensities are different, and a high value means the probabilities are the same. They used the entropy of each of the equal-size blocks of a face image to determine the number of sub-blocks within each block.

### 2.1.3 Classification

Frame-based expression classification does not use temporal information for the input images. It uses the information of current input image. The input image can be a static image or single frame of a video sequence that is treated independently. A wide range of classification methods has been proposed in the literature such as rule-based classifier [138], neural networks (NN) [139; 84], Hidden Markov Model (HMM) [140; 141], support vector machines (SVM) [142; 143], Nearest neighbour (NN) [75], and Bayes classifier [144]. Rule-based classifiers develop certain rules from human point of view, and classify the feature extracted from face image into corresponding classes. Rule-based method can describe facial expression more accurately, and can facilitate the synthesis of facial expression. Neural Networks has wide applications on static image recognition. Gueorguieva use multi-layer perception NN to recognise facial expression [145], which train and test four types networks. [139] extract feature using 2D discrete cosine transform from full face image. One of the limitations of NN based methods is that it is difficult to train classifiers when recognising many unlimited facial expression. The underlying assumption of the HMM is that patterns can be characterised as a parametric random process and that the parameters of this process can be estimated in a precise, well-defined manner. In developing an HMM for a pattern recognition problem, a number of hidden states need to be decided first to form a model. Then, one can train HMM to learn the transitional probability between states from the examples where each example is represented as a sequence of observations. The goal of training an HMM is to maximize the probability of observing the training data by adjusting the parameters in an HMM model with the standard Viterbi segmentation method and Baum-Welch algorithms [140]. After the HMM has been trained, the output probability of an observation determines the class to which it belongs. HMMs have been applied to both face recognition and localization. Samaria [141] showed that the states of the HMM he trained corresponds to facial regions. In [146], a Hidden markov model (HMM) in combination with GentleBoost classifier are used for classifying AUs and their temporal segments. Support vector machine (SVM) based classifiers are pop-

ular in computer vision, which has already been widely applied to facial expression recognition [89; 147]. The main idea behind the SVM is to maximise the distance in the input space between the two classes of data. SVM has two advantages: (1) its ability to work with high-dimensional data; and (2) a high generalisation performance without the need of a priori knowledge, even when the dimension of the input space is very large. The linear, polynomial, and RBF kernels have been shown in [89] to achieve good performance in facial expression recognition. Template based match methods are used in [87; 89] to conduct face and facial expression recognition using LBP based feature, where a template is computed over training image for each class and testing image, and the closet template to the template of testing image is selected using certain similarity measure. Chi square statistic ( $\chi$ ) is a good similarity measure for histogram which is defined as follows:

$$\chi^2(S, M) = \sum_i \frac{(S_i - M_i)^2}{S_i + M_i} \quad (2.4)$$

Nearest neighbour (NN) based classifiers are one of most simple classifiers, which are instance-based and non-parametric methods. In the process of classification, a testing multidimensional feature is assigned the label which has the most votes among the  $k$  training samples nearest to the testing feature. The standard Euclidean distance are a commonly used distance for measuring. NN classifier with Euclidean metric was used for classifying facial expression in [75].

Bayes classifier are based on Bayes theorem, which is a statistical classifier. The naive Bayes (NB) classifier combines the naive Bayes probability model with a decision rule, and use one common rule known as maximum a posteriori decision. Given a set of samples  $X = (X_1, X_2, \dots, X_k)$  with labels  $C = (C_1, C_2, \dots, C_k)$ , the NB classifier assign the classes label  $\hat{y} = C_k$  for  $k$  samples using the following function:

$$\hat{y} = \operatorname{argmax}_{C_k \in C} p(C_k) \prod_{i=1}^n p(x_i | C_k). \quad (2.5)$$

where  $\hat{y}$  denotes the predicted label. Cohn et al. [144] use Bayesian network for facial expression recognition and analyse some Bayes network methods (i.e. Gaussian naive Bayes (NB-Gaussian), Cauchy naive Bayes (NB-Cauhy), and tree-augmented-naive Bayes (TAN), where TAN achieved a best recognition performance (73.2%) on CK dataset for person-independent testing.

### 2.1.4 Challenges

Although different proposed methods for facial expression recognition have achieved good results, there still remains different problems that need to be addressed by the research community. Generally, most existing methods for facial expression recognition are computationally expensive and usually require dimensionally large feature vectors for the classification task. This explains their inability for real-time applications, although they produce good results on different datasets.

Smart meeting, video conferencing and visual surveillance are some of the real world applications that require facial expression recognition system to perform adequately on low resolution images. There exist numerous methods for facial expression recognition but very few of those perform adequately on low resolution images.

More research effort is required for recognizing more complex facial expressions than the six classical expressions such as fatigue and pain, and mental states such as agreeing, disagreeing, lie, frustration and thinking as they have numerous potential application areas.

Other problems include expression intensity estimation, spontaneous expression recognition, micro expression recognition (i.e., brief, involuntary facial expression lasting only 1/25 to 1/15 of a second), misalignment problem, illumination, and face pose variation.

## 2.2 Datasets for facial expression recognition

Most affect recognisers are validated on posed datasets, which differ from naturalistic datasets in terms of illumination conditions, head-pose variations and nature of expressions (subtle vs. exaggerated). The CK [148] and MMI [149] datasets are widely used posed datasets and include basic emotion as well as AU annotations. The enhanced CK dataset [94] provided frame-by-frame AU intensity annotations for the whole CK dataset for 14 AUs and also modified some of the intensity labels that were provided in CK. The CK+ dataset [94] extended CK with spontaneous recordings and novel subjects, annotations and labels (including a non-basic emotion, contempt). A large part of MMI is annotated with temporal segments (neutral, onset, apex, offset).

The Extended CK+ dataset (CK+) [94] is the most widely used data for evaluating facial expression recognition methods, and is publicly available. This dataset contains 593 image sequences of seven basic facial expressions (namely Anger, Contempt, Disgust, Fear, Happiness, Sadness and Surprise). These expressions were



## 2.2. DATASETS FOR FACIAL EXPRESSION RECOGNITION

---

performed by 120 subjects. The age of the participants ranges from 18 to 30 years, 65% of them are female, 81% are Euro-American, 13% are Afro-American, and 6% of other racial groups. Each frame of the image sequences is  $640 \times 480$  or  $640 \times 490$  pixels with an 8-bit grey scale. The video sequences vary in duration (i.e., 10 to 60 frames) and incorporate the onset (i.e., the neutral frame) to peak phase of the facial expression. We used 327 image sequences of seven expressions, where we replaced the expression neutral with Contempt. The top row of Figure 2.7 shows sample images of a subject expressing six expressions. Table 5.8 shows the occurrences of the various expression classes in CK+ dataset.

The MMI dataset [149] is another well-known dataset, which comprises video sequences including both posed and spontaneous expressions. These expressions were performed by 19 subjects (44% female), with age ranging from 19 to 62, and of European, Asian and South American ethnicity. The subjects performed 79 expressions including the six basic facial expressions with neutral frame at the start of each sequence. Every video frame is at  $720 \times 576$  spatial resolution. We converted the original frames into 8-bit greyscale images for our experiments, and extracted the sub-sequence from the neutral frame to the peak phase. In our experiments, 203 image sequences labelled as one of the six basic facial expressions are selected from the MMI dataset. The bottom row of Figure 2.6 shows sample images of the six basic expressions.

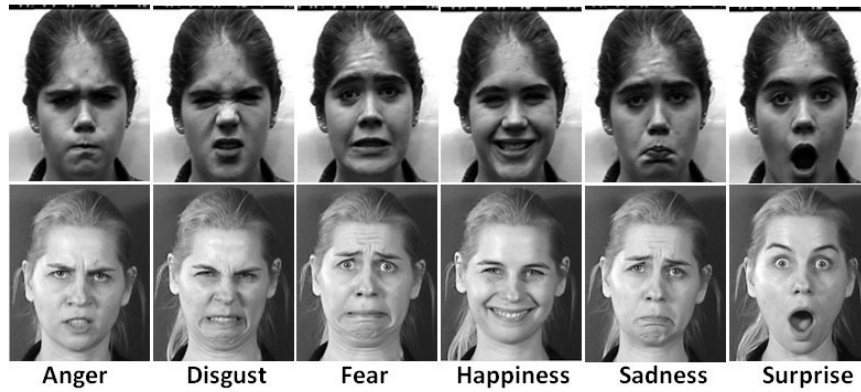


Figure 2.6: Example of expression for the six basic emotions (Left-to-right from top row: anger, disgust, fear, happiness, sadness, and surprise).

Acted facial expressions in the wild (AFEW) and is a dynamic temporal facial expressions data which consists of close to real world environment extracted from movies [150], which comprises 957 video sequences labelled with six prototypic and neutral expressions. The age of subjects ranges from 1 to 70 years, and each clip has been annotated with attributes such as name, age, pose, gender, expression of

### 2.3. MULTI-MODAL EXPRESSION ANALYSIS

---

Table 2.1: Number of image sequences (subjects) for each expression in the CK+ dataset.

Expression	CK+	MMI
Anger	45	32
Disgust	59	28
Fear	25	28
Happiness	69	42
Sadness	28	32
Surprise	83	41
Contempt	18	0
Total	327	203

person and the overall clip expression. The AFEW dataset is collected from close to real environment (actors in movies which is similar to those in the real world), which to some extent addresses the problem existing in current datasets that is usually collected in controlled lab environment. They provided much more available samples in more complex conditions. Static facial expression in the wild (SFEW) has been collected by selecting frames from AFEW [151]. There are 700 facial expression images which are displayed by 95 subjects, which are also labelled with six basic and neutral expressions. Same to AFEW dataset, the SFEW dataset has different head poses, age range, partly occlusions, and the varied illumination. Both AFEW and SFEW provide an useful way to evaluate the facial expression system in varied realistic conditions.

There exist non-posed datasets for several affect recognition contexts including categorical basic/non-basic emotion recognition, AU detection, pain detection and dimensional affect recognition. The GEMEP [152] dataset is collected from professional actor portrayals, and includes 12 non-basic emotions and 6 basic emotions. A subset of this database was used in the FERA challenge. Spontaneous AUs can be studied on the public DISFA [153] dataset as well as the partly public M3 (formerly RU-FACS) [154]. Frame-by-frame AU intensities are provided with DISFA.

## 2.3 Multi-modal expression analysis

Multi-modal expression analysis is briefly introduced in this section.

As one of several forms of non-verbal communication, facial expression might be associated with other forms of non-verbal communication (e.g. gesture) as well as verbal communication [155; 156; 157; 26]. In [26], Cohn et al. explored the relation between facial AUs and vocal prosody for depression detection, which achieved the

same recognition rate (79%) by using AUs and vocal prosody respectively. Gunes and Piccardi [155] combined facial actions and body gestures for 9 expression recognition. They found that recognition from fused face and body modalities performs better than that from the face or the body modality alone.

For facial feature extraction in [156], following frame-by-frame face detection, a combination of appearance (e.g., wrinkles) and geometric features (e.g., feature points) is extracted from the face videos. A reference frame with neutral expression is employed for feature comparison. For body feature extraction and tracking, they detected and tracked head, shoulders and hands by using mean-shift method from the body videos. 2.7 shows examples of the face and body feature extraction in [156]. A total of 152 features for face modality and 170 features for body modality were used for the detection of face and body temporal segments with various classifiers including both frame-based and sequence-based methods. They tested the system on FABO database [155] and achieved recognition rate at 35.22% by only using face features and 76.87% by only using body features. The recognition rate increased to 85% with combination of both face and body features.

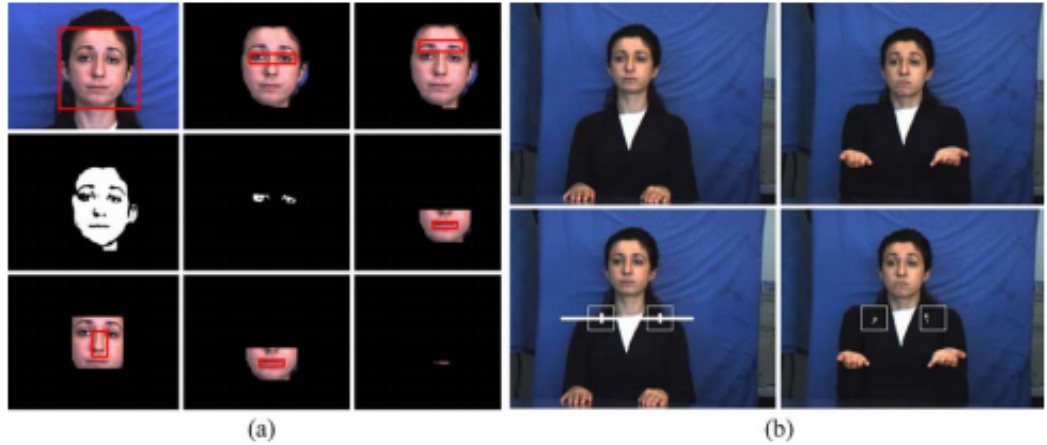


Figure 2.7: Example of the face and body feature extraction employed in the FABO system. (a) Face features. (b) Body features shoulder extraction procedure. Shoulder regions found and marked on the neutral frame (first row), estimating the movement within the shoulder regions using optical flow (second row)

## 2.4 Problem Space for facial expression analysis

Facial expression systems might vary significantly due to their potential applications and aims. There are three important factors which affect the variance of systems: posed or spontaneous facial expression, the illumination condition under

which expressions are performed (strictly under-control conditions or complex real life environment) and individual difference in subjects.

### 2.4.1 Posed vs. spontaneous expression

The expressions can be performed by subjects upon requested or command (e.g. look surprised or look said) or occur spontaneously in natural human-to-human or human-to-computer interaction and communication. How to distinguish deliberately displayed and spontaneous facial expression clearly has drawn more and more attentions, which are very important in daily life communications, security environment, health-care field, and so on. Some psychological research [158; 159] has shown that posed and spontaneous expression are different in timing of display, appearance, movement of head pose and some other body gestures. Most of existing methods on distinguish posed and spontaneous expression on the area of computer vision focus on analysing these two on the visible image domain (e.g. geometric information). A few geometric descriptors extracted from several facial regions are used for distinguishing, such as the movements of eye brows and eyes [160]. Temporal information and facial action is also used for analysing posed and spontaneous expressions in [161], which shows spontaneous smiles have smaller amplitude and more consistent relation between amplitude and duration than posed one. Compared with posed expression, the spontaneous one are much more complicated to be recognised in recognition systems since spontaneous expressions normally have lower intensity (i.e. smaller and slower movements) and display along with the changes of head pose and speech action. As an instance, when FACS action units which are trained on Cohn-Kanade (CK+)[94] dataset of posed facial expression are applied to a dataset of spontaneous facial expression (RU-FACS dataset [154]) for testing, the recognition rate drop by over 20% [160]. Therefore, how to recognise spontaneous expression more accurately and effectively in real settings is one of future research direction, which is required to pay more attention.

### 2.4.2 Under-control condition vs. real life

Most of existing datasets are collected in controlled laboratory conditions (i.e. lighting, camera position, frame rate, etc.) which do not reflect the type of conditions seen in real environment [162]. However, it is not good to focussing too much on the laboratory environment for the field of facial expression recognition. This is because the extracted feature trained in the dataset obtained under laboratory conditions might not be applicable in real life applications with more complex conditions, leading to the large decrease of recognition performance. In general, recognising facial

expression in real environment is a much more challenging problem. For instance, a smile detector based on linear regression performed proposed in [163] achieves a high recognition rate of 97% using CK+ dataset for experiments, while decrease to only 72% when applied to natural environment.

### 2.4.3 Individual difference in subjects

Face shape, texture, colour, and hair vary with ethnic background, sex and age [164; 165]. For instance, infants have smoother and less texture skin, and lack hair in the brows. The eye opening and contrast between iris differ significantly between Asians and Northern Europeans, which may affect the robustness of eye tracking and facial features analysis more generally. Facial parts may be occluded by beards, glasses, or jewellery. Such individual differences in appearance may have important consequences for recognition. Few attempts to study their influence have been made. An exception was a research by Zlochower et al.[165], who found that optical flow based method and high gradient component detection that are optimised for young adults performed less well when used in infants. Some features (e.g., reduced texture, increased fatty tissue, lack of transient furrows, etc.) of infants skin may all have contributed to the differences observed in face analysis between infants and adults. In addition to individual differences in appearance, there are individual differences in expressiveness, which refers to the degree of facial plasticity, morphology, frequency of intense expression, and overall rate of expression. Individual differences in these characteristics are well established and are an important aspect of individual identity [166] (these individual differences in expressiveness and in biases for particular facial actions are so strong that they may be used as a biometric to augment the accuracy of face recognition algorithms [166]). An extreme example of variability in expressiveness occurs in individuals who have incurred damage either to the facial nerve or central nervous system [167; 168]. To develop algorithms that are robust to individual differences in facial features and behaviour, it is essential to include a large sample of varying ethnic background, age, and sex, which includes people who have facial hair and wear jewellery or eyeglasses and both normal and clinically impaired individuals.

### 2.4.4 Transitions among expressions

A simplifying assumption in facial expression analysis is that expressions are singular, starting and ending with a neutral position. In the real environment, facial expression is more complex, especially at the level of action units. Action units may occur in combinations or show serial dependence. Transitions from action units or

combination of actions to another may involve no intervening neutral state. Parsing the stream of behaviour is an essential requirement of a robust facial analysis system, and training data are needed that include dynamic combinations of action units, which may be either additive or non-additive. As shown in Figure 1.3, an example of an additive combination is smiling (AU 12) with mouth open, which would be coded as AU 12+25, AU 12+26, or AU 12+27 depending on the degree of lip parting and how far the mandible was lowered. In the case of AU 12+27, for instance, the facial analysis system would need to detect transitions among all three levels of mouth opening while continuing to recognize AU 12, which may be simultaneously changing in intensity.

Non-additive combinations represent further complexity. Following usage in speech science, we refer to these interactions as co-articulation effects. An example is the combination AU 12+15, which often occurs during embarrassment. Although AU 12 raises the cheeks and lip corners, its action on the lip corners is modified by the downward action of AU 15. The resulting appearance change is highly dependent on timing. The downward action of the lip corners may occur simultaneously or sequentially. The latter appears to be more common [83]. To be comprehensive, a database should include individual action units and both additive and non-additive combinations, especially those that involve co-articulation effects. A classifier trained only on single action units may perform poorly for combinations in which co-articulation effects occur.

## Chapter 3

# Patch based facial expression recognition framework

A framework based on sparse representation which is referred as Framework 1 in this chapter is proposed for recognising facial expression. In Framework 1, some prominent facial patches which depends on the location of facial landmarks are extracted during emotion classification. These active patches are then selected and processed to obtain the salient patches which contain discriminative features for classification of each pair of expressions, as different patches have various contributions on different pairs of expression classes. Classifiers using one-against-one strategy are employed to classify these features. An improved sparse representation technique is applied to extract sparse features, which achieves promising performance.

### 3.1 Introduction

In comparison with extracting facial features from the whole image, patch based methods which extract facial features by dividing a face image into several sub-regions (patches) have shown promising performance [114; 115; 116; 117] (see the detailed report in Section 2.1.2). Most of the research for facial expression recognition are conducted on different datasets with suitable performance criteria befitting to the dataset. For instance, the selection of prominent facial areas improves the performance. However, in most of the literature, the size and position of these facial patches show various contributions on different datasets, and it is difficult to conceive a generic system using these methods. Therefore, to address these limitations, Framework 1 attempts to identify the salient facial areas by generalising discriminative features for expression classification. The selection of salient patches retaining discriminating features between each pair of facial expressions improves

### 3.2. PROPOSED FRAMEWORK

---

the recognition accuracy. The size and location of patches are remain same for different database.

Recently, sparse representation for facial expression recognition was proposed in the Wright’s paper [108] based on principles of compressed sensing [169]. However, their methods still use full face image which might not be able to handle occlusion. The proposed work in this chapter shift the focus to part-based representation which can deal with occlusion, and improve the robustness of the recognition. Inspired by Wright’s work [108], Framework 1 attempts using sparse representation to extract feature on divided patches to improve the recognition performance, and analyses how different patches contribute different facial expression.

There are three main contributions in Framework 1: (a) part-based method which is robust to image alignment; (b) analysis on the contribution to expressions of different facial patches and; and (c) sparse representation which reduces the dimensionality of features, improves the semantic meaning and represents the face image more efficiently.

## 3.2 Proposed framework

Changes in facial expressions involve contraction and expansion of facial muscles which alter the position of facial landmarks. Along with the facial muscles, the texture of facial parts also changes. How the changes of these facial parts could contribute the recognition performance is still to be addressed. Thus, this chapter attempts to understand the contributions of different facial areas for automatic expression recognition. In other words, this chapter explores facial patches which generate discriminative features to separate two expressions effectively.

The overview of Framework 1 is shown in Figure 3.1. The accuracy of landmarks detection and extraction of appearance features from active face regions improve the performance of expression recognition [36]. Therefore, the first step is to localise the face followed by detection of the landmarks. A learning-free approach is proposed in which the eyes and nose are detected in the face image and a coarse region of interest (ROI) is marked around each of these facial organs. The lip and eyebrow corners are detected from their ROIs. Locations of active patches are defined with respect to the location of facial landmarks. Figure 3.2 illustrates the steps involved in facial landmark detection and active patch extraction. In the training stage, all the active facial patches are evaluated and the ones with features of maximum variation between pairs of expressions are selected. These selected features are projected onto a lower dimensional subspace and classified into different expressions using a multi-class classifier. The training phase includes pre-processing, selection



### 3.3. FACIAL LANDMARK DETECTION

---

of facial patches, extraction of features for sparse representation, and learning of the multi-class classifiers. In an unseen image, the process first detects the facial landmarks, then extracts the features from the selected salient patches, and finally classifies the expressions.

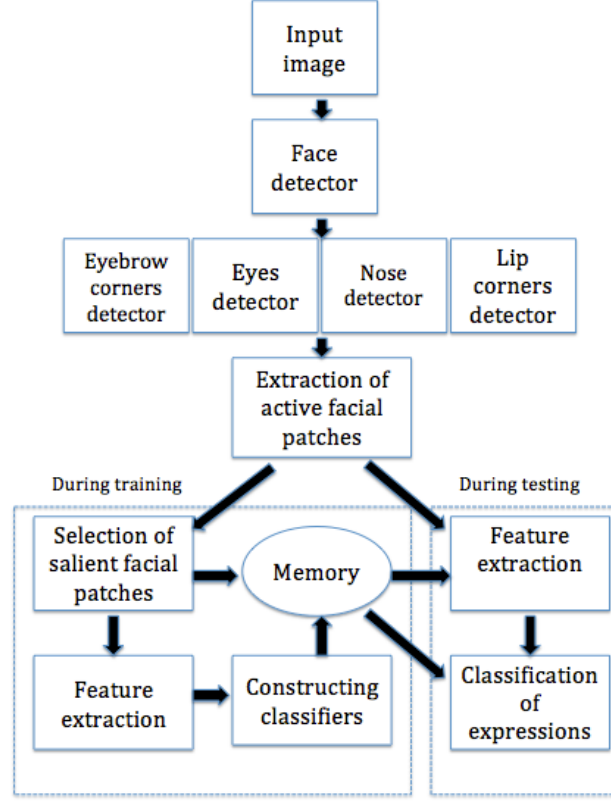


Figure 3.1: Framework 1 for facial expression recognition.

### 3.3 Facial landmark detection

The facial patches which are active during different facial expressions have been reported in [119]. More specifically, some facial patches make contributions and are active when all basic expressions happen, while some are active only to a single expression. The results indicate that common patches which are active are lying below the eyes, in between the eyebrows, around the nose and mouth corners. To extract these patches from face image, we need to locate the facial components first followed by the extraction of the patches around these organs. Unzueta et al. [170] proposed a robust, learning-free, lightweight generic face model fitting method for localisation of facial organs. Using local gradient analysis, this method finds the

### 3.3. FACIAL LANDMARK DETECTION

---

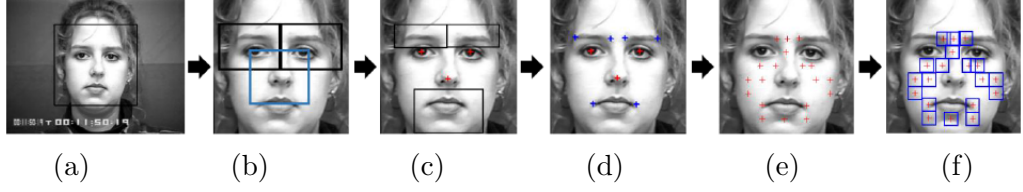


Figure 3.2: Framework for automated facial landmark detection and active patch extraction: (a) face detection, (b) coarse ROI selection for eyes and nose, (c) eyes and nose detection followed by coarse ROI selection for eyebrows and lips, (d) detection of corners of lip and eyebrows, (e) finding the facial landmark locations, and (f) extraction of active facial patches.

facial features and adjusts the deformable 3D face model so that its projection on image will match the facial feature points. Such a learning-free approach is adopted for localisation of facial landmarks in Framework 1. The active facial patches with respect to the position of eyes, eyebrows, nose, and lip corners are extracted using the geometrical statistics of the face.

#### 3.3.1 Pre-processing

A  $3 \times 3$  Gaussian mask (low pass filtering) is employed to remove some noise from a face image, which is followed by a popular face detector for face localisation. Viola-Jones face detector [57] is used in this Chapter due to its lower computational complexity and its detection accuracy under near frontal and upright pose face image, where Viola-Jones face detector use Haar-like features and Adaboost learning technique. To remove the influence of scale and position, integral image is used for detection. The localised face is then scale to normal resolution, which can make the proposed method shift-invariant (insensitive to the location of the face in the face image). To improve the image contrast, histogram equalisation is applied for illumination correction.

#### 3.3.2 Eye and Nose Localisation

To reduce the computational complexity as well as the false detection rate, the coarse ROI for eyes and nose are selected according to geometrical structure of face (i.e., the ROI for nose is in the centre of the detected face and the two eyes are located symmetrically at the top one-third position of the face). Both eyes are detected separately using Haar classifier eye detector for each eye. The Haar classifier returns the vertices of rectangular area of detected eyes. The eye centres are computed as the mean of these coordinates. Similarly, the nose position is also

detected using the Haar cascades. In the case where the eyes or nose are not detected by the Haar classifiers, Framework 1 relies on the landmark coordinates detected by anthropometric statistics of face. The position of eyes are used for up-right face alignment since the positions of eyes do not change with facial expressions.

#### 3.3.3 Lip corner detection

Inspired by the work of Nguyen et al. [171], Framework 1 uses facial topographies for detection of lip and eyebrow corners. The ROIs for lips and eyebrows are selected as a function of face width positioned with respect to the facial organs.

The ROI for mouth is extracted using the position of nose as reference. The upper lip always produces a distinct edge which can be detected using a horizontal edge detector. Sobel edge detector [172] is used for this purpose due to its good performance on edge detection. In images with different expressions, numerous spurious edges are extracted, which is then removed by using the classical and popular segmentation method-Otsu thresholding method [173]. In this process, a binary image is obtained which contains many connected regions. Using connected component analysis [172], the spurious components having an area less than a threshold are removed. Morphological dilation operation is then applied on the resulting binary image. Finally, the connected component with the largest area just below the nose region is selected as the upper lip region. Figure 3.3 shows different stages of the process. The algorithm is as follows:

Algorithm. Lip corner detection

Given: Aligned face ROIs and nose position.

1. Select coarse lips ROI using face width and nose position.
2. Apply Gaussian blur to the lips ROI to remove the noise of image.
3. Apply horizontal Sobel operator for edge detection
4. Apply Otsu thresholding.
5. Apply morphological dilation operation.
6. Find the connected components.
7. Remove any spurious connected components by thresholding.
8. Scan image from the top and select the first connected component as upper lip position.
9. Locate the left- and right-most positions of connected component as lip corners.

### 3.3. FACIAL LANDMARK DETECTION

---

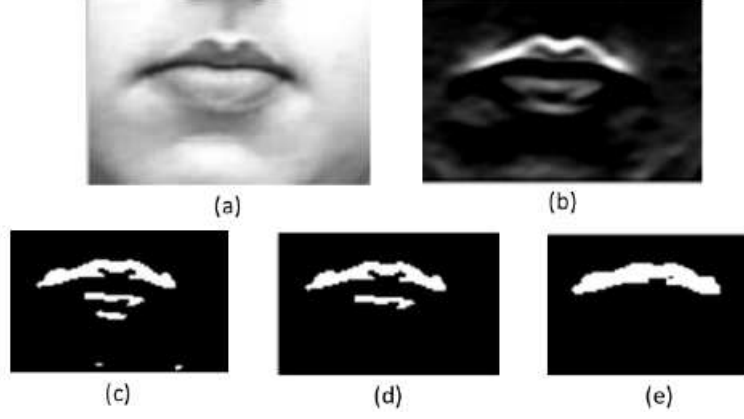


Figure 3.3: Lip corner localisation: (a) lips ROI, (b) applying horizontal Sobel edge detector, (c) applying Otsu thresholding, (d) removing spurious edges, and (e) applying morphological operations to render final connected component for lip corner localisation.

Sometimes, because of shadow below the nose, the upper lip could not be segmented properly which is shown in Figure 3.4. As shown in Figure 3.4, the upper lip is not segmented as a whole and the connected component obtained at the end resemble half of the upper lip. Hence, the extreme ends of this connected component does not satisfy the bilateral symmetry property, i.e., the lip corners should have been at more or less equal distances from the vertical central line of a face. These situations are detected by applying a threshold to the ratio of distance between the lip corners to the maximum of distances of the lip corners from the vertical central line. In such a case, the second connected component below the nose is considered as the other part of the upper lip. Thus the lip corners are detected with the help of two connected components. By using the above methods, false detection of lip corner points is minimised.

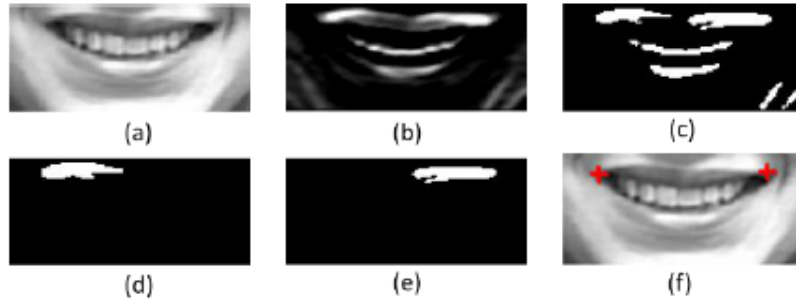


Figure 3.4: Lip corner localisation where the upper lip is not entirely connected, (a-c) same as Figure 3.3, (d-e) selection of two connected components by scanning from top, and (f) localized lip corners.

#### 3.3.4 Eyebrow corner detection

With the knowledge of positions of the eyes, the coarse ROIs of eyebrows are selected. The eyebrows are detected following the same steps as that of upper lip detection. As can be observed in the experiments that performing an adaptive threshold operation before applying horizontal Sobel operator improves the accuracy of eyebrow corner localisation. The use of horizontal edge detector reduces the false detection of eyebrow positions due to partial occlusion by hair. The inner eyebrow corners are detected accurately in most of the images. Figure 3.8 illustrates the intermediate steps in eyebrow corner detection. The algorithm is as follows.

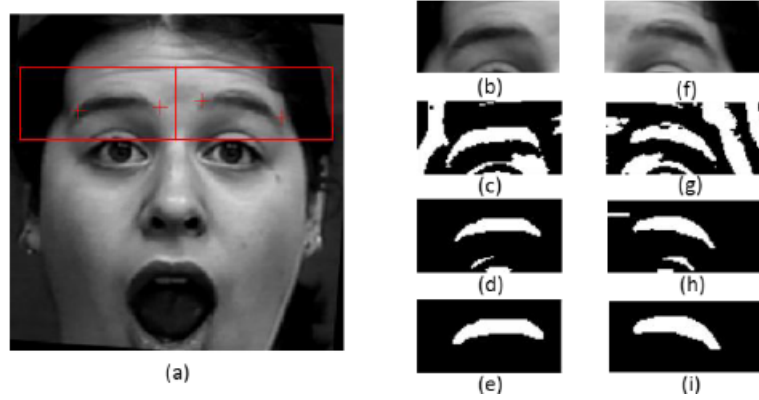


Figure 3.5: Eyebrow corner localisation: (a) rectangles enclosing ROI and plus marks showing the detection result, (b and f) eye ROIs, (c and g) applying adaptive threshold on ROIs, (d and h) applying horizontal Sobel edge detector followed by Otsu thresholding and morphological operations, and (e and i) final connected components for corner localisation.

Algorithm. Eyebrow corner detection

Given: localised eye positions

1. Select eyebrow regions using eye positions.
2. Apply Gaussian blur to the eyebrow ROI.
3. Apply adaptive Otsu-thresholding.
4. Apply horizontal Sobel operator for edge detection.
5. Apply Otsu-thresholding.
6. Apply morphological dilation operation.
7. Find the connected components.
8. Remove spurious connected components by thresholding.

9. Scan image from the right and select the first and second connected components as left and right eyebrow positions.
10. Locate the left and right most positions of connected component as eyebrow corners of left and right eyebrows.

## 3.4 Extraction of active facial patches

During an expression, the local patches are extracted from the face image depending on the position of the active facial muscles. The appearance of facial regions exhibiting considerable variations during one expression is considered. For example, wrinkles in upper nose region are prominent in disgust expression and absent in other expressions. Similarly, regions around lip corners undergo significant changes and their appearance features are dissimilar for different expressions. The active facial patches are shown in Figure 3.6.

The patches do not fixed positions on the face image. However, the locations depend on the positions of facial landmarks. The size of all facial patches are kept equal and is determined experimentally to be approximately one-ninth of the width of the face. The patches are referred by the numbers assigned to them. As illustrated in Figure 3.6,  $p_1$ ,  $p_4$ ,  $p_{18}$ , and  $p_{19}$  are directly extracted from the position of lip corners and inner eyebrows receptively,  $p_{16}$  is at the centre between both eyes; and  $p_{17}$  is the patch above  $p_{16}$ .  $p_3$  and  $p_6$  are located midway between the nose and left eye and right eye, respectively.  $p_{14}$  and  $p_{15}$  are respectively located just below the left and right eyes.  $p_2$ ,  $p_7$  and  $p_8$  are clubbed together and located to the left of the nose.  $p_9$  is located just below  $p_1$ . Similarly,  $p_5$ ,  $p_{11}$ ,  $p_{12}$  and  $p_{13}$  are located to the right of the nose.  $p_{10}$  is located at midway between  $p_9$  and  $p_{11}$ . The algorithm of selecting these patches is as follows.

Algorithm. Selection of patches

1. Select  $p_1$ ,  $p_4$ ,  $p_{18}$ , and  $p_{19}$  directly from the locations of lip corners and inner eyebrows, respectively.
2. Select  $p_9$  and  $p_{11}$  respectively below  $p_1$  and  $p_4$ .
3. Select  $p_{16}$  using the middle point between inner corners of the two eyes.
4. Select  $p_{17}$  just above  $p_{16}$ .
5. Select  $p_3$  and  $p_6$  using the location of nose and inner corners of the two eyes.
6. Select  $p_{14}$  and  $p_{15}$  just below the connect line between corners of each eye.
7. Select  $p_2$ ,  $p_7$ ,  $p_8$ ,  $p_5$ ,  $p_{12}$  and  $p_{13}$  using the nose location.
8. Select  $p_{10}$  which is located at the mid position between  $p_9$  and  $p_{11}$ .

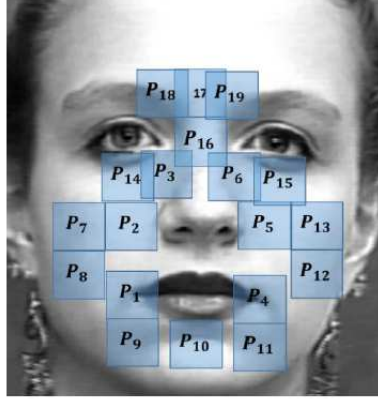


Figure 3.6: Position of facial patches.

### 3.5 Feature extraction and classification

#### 3.5.1 Sparse representation

Let a set of  $N$  training facial images (different selected patches in Framework 1) be separated into  $K$  different facial identity classes. Each of the facial images is scanned row-wise in order to form an  $F$ -dimensional vector. Dimensionality reduction is then applied, like Principal Component Analysis (PCA) [174], in order to form from each image a vector  $x_i \in R^f$ , with  $f \ll F$ , which is normalised in such a way that  $\|x_i\| = 1$ . Let the dictionary (matrix)  $X$  be defined as  $X = [x_1 | \dots | x_N] \in R^{fN}$ . Let a test image  $y \in R^f$ . In [108], a method for feature extraction via sparse decomposition to achieve face recognition is proposed. That is, it motivated the use of an  $l_1$  optimization problem in Framework 1 to find a sparse vector of weights  $w$  which depicts the contribution of each facial training image  $x_i$  in the formation of the test facial image  $y$ . According to [108], let  $\varepsilon$  be a threshold, the optimization problem for finding the sparse vector  $w$  is then

$$\tilde{w} = \arg \min \|w\|_1 \quad \text{subject to} \quad \|Xw - y\|_2^2 < \varepsilon \quad (3.1)$$

After determining the optimal vectors  $\tilde{w}$  according to (4.13), the method attempts to classify image  $y$  to one of the  $K$  facial identity classes. Let  $\delta_k(\tilde{w})$  be a new vector whose only non-zero entries are the entries in  $\tilde{w}$  that are associated with class  $k$ . Using only the coefficients associated with the  $k$ th facial identity class, the given test sample  $y$  can be approximated as  $\tilde{y}_k = X_k(\tilde{w})$ . Image  $y$  is classified based on these approximations to the object class that minimizes the residual between  $y$

### 3.5. FEATURE EXTRACTION AND CLASSIFICATION

---

and  $y^k$  [108], i.e.,

$$l(y) = \arg \min_k r_k(y) = \|y - \tilde{y}_k\|_2. \quad (3.2)$$

Moreover, in order to model pixel corruptions and deal with the presence of occlusion, an error vector  $e$  has been taken into consideration in the optimisation problem. Assuming that the error vector  $e$  has sparse non-zero entries with respect to the natural pixel coordinates, the dictionary can be changed to

$$X_{(e)} = [X, I] \in R^{f \times (n+f)}, \quad (3.3)$$

where  $I$  is the identity matrix. The vector  $w_e = \begin{bmatrix} w \\ e \end{bmatrix}$  can then be seek from the optimization of

$$\tilde{w}_{(e)} = \arg \min \|w_e\|_1 \quad \text{subject to} \quad \|X_e w_e - y\| < \varepsilon. \quad (3.4)$$

To solve the optimisation problem (4.13) and (3.4), the  $l_1$ -magic software package [175] is used in Framework 1.

Once the sparse solution  $\tilde{w}_e$  is computed from (3.4), the facial image cleaned from corruption or occlusion,  $y_r = y - \tilde{e}$ , is used for classification using the rule:

$$l(y) = \arg \max_k r_k(y) = \|y_r - X\delta_k(\tilde{w})\|_2. \quad (3.5)$$

If the database comprises  $K$  facial identity classes, vector  $\tilde{w}$  should contain high valued coefficients that correspond to the facial identity class to which image  $y$  belongs, and very low (or probably zero) values for all other images. Take Figure 3.7(a) for example, the test image (to the left of the vertical 0 axis) is represented by the gallery (training) images (the three images to the right of the vertical 0 axis which are selected from the training images). For face recognition, if the test image is person A, the coefficients that correspond to person A in the training images should contain high value (i.e., the high responses in the plot) and the coefficients that correspond to persons should contain low values (i.e., the low responses in the plot). Using this method for facial expression recognition, the sparse representation should have high valued responses for the expressive images of the same facial expression class of the test facial image and low valued responses for all other classes. The image that is decomposed in Figure 3.7(b) is an expressive facial image of disgust. As it can be seen in the decomposition, high responses for images of the same person that depict a different facial expression are achieved. Removing the images of the same person from the gallery is also tested in the experiments. In Figure 3.7(c)



the decomposition of the same person from Figure 3.7(b) is depicted but now the gallery does not contain images of the same person and of the neutral expression. As it can be seen, the decomposition is not sparse and does not produce high valued responses for the coefficients that correspond to the correct facial expression class (i.e., it fails to achieve decomposition using images from the same expression class). In Figure 3.7, x axis describes sample images from training set, and y axis describes the values of sparse coding.

Therefore, the direct use of the expression image results in a rather difficult decomposition in terms of facial expression classes using the method of [108]. This is due to the fact that the features of the same facial identity influence the result to a greater extent than the features of the same facial expression class. It is clear that in order to use such an algorithm it is necessary to eliminate the contribution of the facial identity in the description of expressions. That is, it is necessary to seek as much as possible person-independent descriptions of expressions. Such representations can be found using the difference images. An example of the decomposition using the difference images is shown in Figure 3.7(d). As it can be seen: 1) the decomposition does not result in high valued responses for difference images of the same person, and 2) the decomposition, using the difference images produces a meaningful sparse representation where high valued responses are obtained for the difference images of the same facial expression class of disgust. In Framework 1, the difference image is used to replace the original image.

#### 3.5.2 Learning salient facial patches across expressions

In most of the related literature, all the facial features are concatenated to recognise the expression. However, this generates a feature vector of high dimension. It is noted that the features from a few facial patches can replace the high dimensional features without significant diminution of the recognition accuracy. From human perception, not all facial patches are responsible for recognition of one expression. The facial patches which contain more movement information and are responsible for recognition of each expression can be used separately to recognise that particular expression. Based on this hypothesis, the performance of each facial patch is evaluated for recognising different expressions.

In addition, some expressions share similar movements of facial muscles, and features of such patches are redundant for classifying the expressions. Therefore, after extracting the active facial patches, the salient facial patches responsible for discrimination between each pair of basic expressions can be selected. A facial patch is considered to be discriminative between two expressions if the features extracted

### 3.5. FEATURE EXTRACTION AND CLASSIFICATION

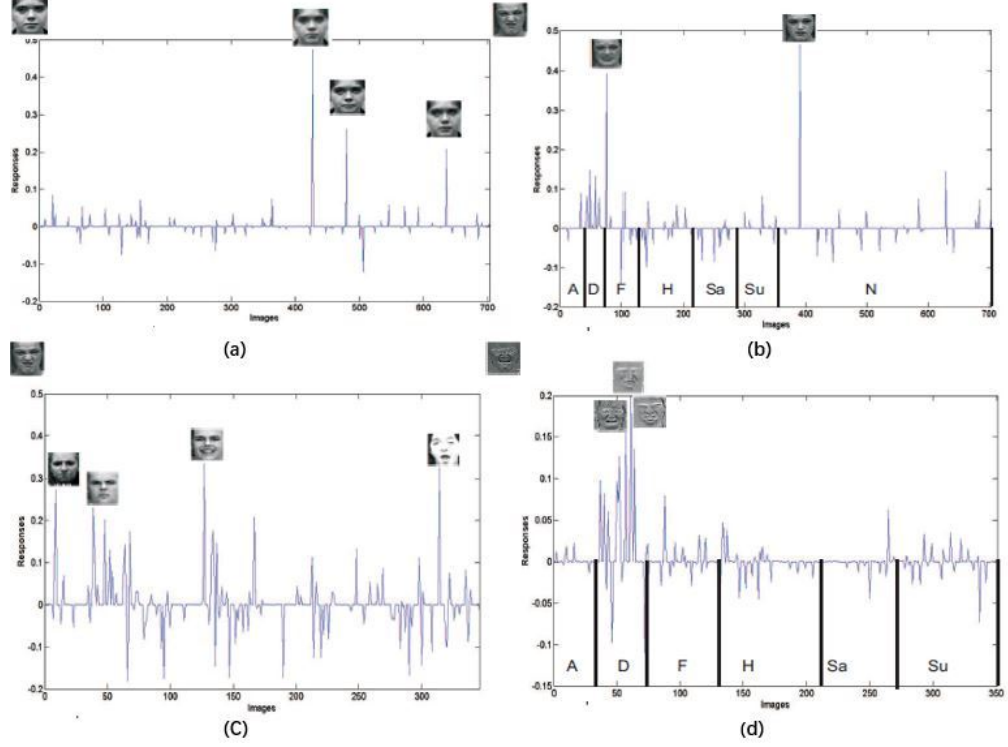


Figure 3.7: (a) Sparse decomposition of a facial image including images of the same person in the dictionary; (b) Sparse decomposition of an expressive facial image including images of the same person in the dictionary; (c) Sparse decomposition of the same image excluding images of the same facial class from the dictionary; and (d) Sparse decomposition of the differences images.

from this patch can classify the two expressions accurately. Note that not all active patches are salient for recognition of all expressions.

The saliency of all facial patches for all pairs of expressions are evaluated in Framework 1, and it is expressed in terms of saliency scores. The saliency of a patch represents the ability of the features from the patch to accurately classify a pair of expressions. The saliency score of a patch between a pair of expressions is the classification accuracy of the features from that patch in classifying two expressions. Saliency score of all patches for each pair of expressions are then calculated.

One-against-one strategy is applied for expression classification. While classifying between a pair of expressions, the sparse features are extracted from those facial patches that have high saliency score. The feature vectors from the salient patches are concatenated to construct a higher dimensional feature vector. Thus, the dimension of the feature vector depends upon the number of patches selected for classification.

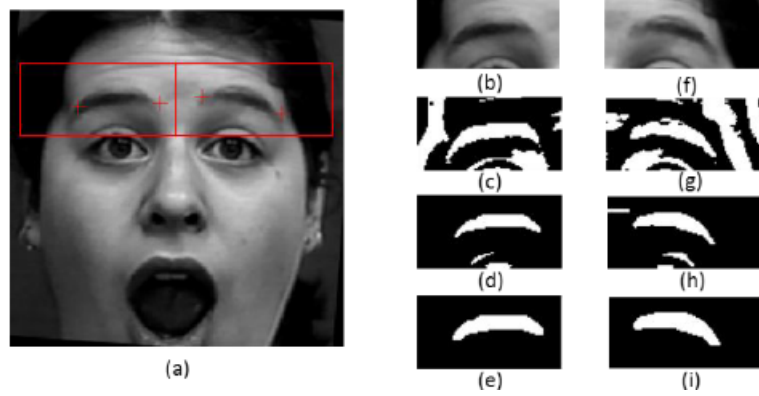


Figure 3.8: Eyebrow corner localisation: (a) rectangles enclosing ROI and plus marks showing the detection result, (b and f) eye ROIs, (c and g) applying adaptive threshold on ROIs, (d and h) applying horizontal Sobel edge detector followed by Otsu thresholding and morphological operations, and (e and i) final connected components for corner localisation.

PCA is applied to reduce the dimensionality of the feature vector. Thus, by projecting the feature vectors from salient patches to the optimal sub-space obtained by above method, the lower dimensional vector with maximum discrimination for different classes can be found. The weight vectors, corresponding to the salient patches of each pair of expression classes, generated during the training stage are used during testing. After reduction of dimensionality using PCA, SVM is used for classification (detailed discussion on SVM can be seen in Section 2.1.3).

## 3.6 Experiments

Framework 1 was evaluated by using two widely used facial expression databases, i.e., Japanese Female Facial Expressions (JAFPE) [176] and Cohn-Kanade (CK+) [94]. Ten-fold cross validation was employed to evaluate the performance of Framework 1. As discussed in Section 3.3, face detection was carried out on all images followed by scaling to bring the face to a common resolution. Facial landmarks were detected and salient facial patches were extracted from each face image. During the training stage, a SVM classifier was trained between each pair of expressions. Here the training data were the concatenated sparse features extracted from the salient patches containing discriminative characteristics between the given pair of expression classes. Similarly,  ${}_6C_2$  numbers of SVM classifiers were constructed and used for evaluating the performance on the test set.

### 3.6.1 Experiments on the Cohn-Kanade database

The Cohn-Kanade database is introduced in Chapter 2. In the experiments, the last image from each sequence was selected where the expression is at its peak intensity. The number of instances for each expression varies according to its availability. In the experiments on CK+ database, Framework 1 used 329 images in total: anger (41), disgust (45), fear (53), happiness (69), sadness (56), and surprise (65). Table 3.1 shows the confusion matrix of six emotions using Framework 1.

Table 3.1: The confusion matrix using Framework 1 on CK+ database.

	A	D	F	H	Sa	Su
Anger(A)	87.8	0	0	0	7.32	4.88
Disgust(D)	0	93.33	0	4.44	0	2.22
Fear(F)	0	1.88	94.33	0	1.88	1.88
Happiness(H)	1.44	2.89	0	94.2	0	1.44
Sadness(Sa)	1.78	0	0	1.78	96.42	0
Surpsie(Su)	0	0	0	1.53	0	98.46

#### 3.6.1.1 Optimum number of salient patches

The number of patches used for classification has an influence on the performance in terms of accuracy and speed. Figure 3.9 shows the average of accuracies of all expressions with respect to the number of salient patches used for classification with a face resolution of  $96 \times 96$ . It is apparent from Figure 3.9 that the use of features from all the 19 patches can classify all expressions with an accuracy. As can be seen from Figure 3.9, the patches of 4, 5, 11, and 13 perform better than the other patches, which shows higher recognition rate.

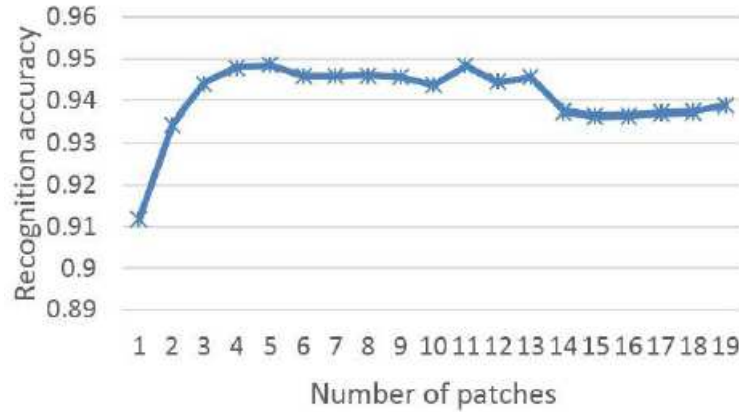


Figure 3.9: The recognition rate using different number of salient patches.

### 3.6. EXPERIMENTS

---

It is clear that even the use of features of a single salient patch can discriminate between each pair expressions efficiently with recognition rate. This implies that the use of rest of the features from other patches contribute minimum towards the discriminative features. The more patches are used, the larger is the size of the feature vector, which increases the computational burden.

Therefore, instead of using all the facial patches, it is reasonable to use some specific salient facial patches for expression recognition. This might improve the computational efficiency as well as robustness of the features (e.g., when a face is partially occluded). Framework 1 selects top four salient patches (i.e., 4, 5, 11, and 13) for the experiments resulting in good accuracy (close to 95%).

#### 3.6.2 Experiments on JAFFE Database

JAFFE database is another classical facial expression database for evaluating the recognition rate of a system. The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The images in the database are well aligned and with small variation in head pose. While testing on the JAFFE database, the same parameters obtained for CK+ dataset were used. In this experiment, 186 images were selected in total: anger (30), disgust (32), fear (29), happiness (31), sadness(31), and surprise (30). The confusion matrix, as in Table 3.2, shows the consistent performance of Framework 1.

Table 3.2: The confusion matrix using Framework 1 on JAFFE database.

	A	D	F	H	Sa	Su
Anger(A)	100	0	0	0	0	0
Disgust(D)	0	93.75	0	0	0	6.25
Fear(F)	6.89	6.89	86.2	0	0	0
Happiness(H)	0	0	0	96.77	0	3.22
Sadness(Sa)	9.67	6.45	0	6.45	77.41	0
Surpsie(Su)	0	3.33	0	0	0	96.66

An overall accuracy of 91.8% was obtained. From Table 3.2, we can see Framework 1 performed worst for sadness expression as it misclassified sadness as anger.

#### 3.6.3 Experiments on fused database

For generalisation, the samples of two databases are combined together to train the classifier [60]. Sample level combination was performed by putting the images of

### 3.7. CONCLUSION

---

both databases together. The training set was constructed by randomly selecting 90% of the data for each expression of each database. The rest data were used as testing set. The models were trained, and their performances were evaluated on samples of individual databases in the testing set. This experiment was repeated for ten times. By learning the features from different databases, the classifier performs better in various situations. All samples were treated with equal probability of selection for training or testing. Therefore, it is expected that the database with more samples should dominate in performance. However, Framework 1 performed well on both databases with a significant accuracy of 89.64% and 85.06% on CK+ and JAFFE databases, respectively. The top four salient patches for classification of each pair of expressions is provided in Table 3.3

Table 3.3: The salient patches derived from fusion CK+ and JAFFE Dataset.

	A	D	F	H	Sa	Su
A	-	P1,P4,P9,P10	P2,P4,P5,P6	P1,P4,P9,P11	P1,P9,P10,P18	P1,P4,P9,P10
D	-	-	P1,P2,P4,P8	P1,P4,P8,P9	P1,P4,P8,P9	P1,P5,P11,P12
F	-	-	-	P1,P2,P4,P8	P1,P9,P18,P2	P1,P2,P5,P6
H	-	-	-	-	P1,P7,P9,P11	P2,P4,P5,P11
Sa	-	-	-	-	-	P1,P9,P10,P11
Su	-	-	-	-	-	-

## 3.7 Conclusion

This chapter has presented a computationally efficient facial expression recognition framework for accurate classification of the six universal expressions. It investigates the relevance of different facial patches in the recognition of different facial expressions. All major active regions on face that are responsible for the face deformation during an expression are extracted. The position and size of these active regions are defined with respect to some facial components. The framework analyses the active patches and determines the salient areas on face where the features are discriminative for different expressions. Using the sparse features from the salient patches, the framework performs the one-against-one classification task and determines the expression based on majority vote.

Expression recognition is carried out using Framework 1. Promising results are obtained by using sparse features of the salient patches. Extensive experiments were carried out on two facial expression databases and the combined dataset. The classification between anger and sadness is found to be troublesome for Framework 1 on both databases. The framework appears to perform well in CK+ dataset with an F-score of 94.39%. Using the salient patches obtained by training on CK+

### 3.7. CONCLUSION

---

dataset, the framework achieves an recognition rate of 91.8% in JAFFE dataset. This demonstrates the efficacy of the proposed framework. The performance of Framework 1 is comparable with earlier similar works, nevertheless Framework 1 is fully automated.

## Chapter 4

# Spatial-Temporal Framework Based on Histogram of Gradient and Optical Flow

This chapter proposes a novel framework referred as Framework 2 for recognising facial expression by using two different types of geometric features, where the classical Lucas-Kanade optical flow [177] method is employed to track the movement of certain facial landmarks, and an extended descriptor based on spatial pyramid histogram of gradient (PHOG) [99] is designed for extracting changes in face shape. These two geometric features are integrated to better represent the dynamic information in terms of individual points and local shape. To analyse the discriminative power of different facial components, Framework 2 extracts the proposed spatio-temporal descriptor on different facial components. A series of experiments are carried out to evaluate the performance of the proposed descriptor, where it is demonstrated that the integrated framework achieves a better performance than using individual descriptor, and also outperforms most of state-of-the-art methods.

### 4.1 Introduction

Facial feature representation is classified into two categories: spatial and spatio-temporal. The former encodes face image sequences frame by frame, while the latter considers an image sequence as spatio-temporal volume using temporal windows which may exploit the relationship between adjacent frames. Most existing methods on facial expression recognition focus on the spatial representation, where they analyse and extract facial features in a single frame of an image sequence, i.e., recog-



dition of static expression. These methods have mainly concentrate on attempting to capture expressions through either action units [34; 178] or via discrete frame extraction techniques [179]. They require either manual selection of facial features in order to determine where the particular changes in the facial region occur, or the subjective thresholding for feature selection, which means that any classification is highly dependent on subjective information in the form of a thresholding or other a priori knowledge.

A facial expression involves a dynamic process, and the dynamic information such as the movement of facial landmarks and the change in facial shape contains useful information that can represent a facial expression more effectively. Thus, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Previous recognition methods on video sequences tend to only focus on the movement of facial landmarks, not analysing the variation of facial shape. Framework 2 utilises two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on PHOG to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. We view an image sequence as a spatio-temporal volume, and use temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. In this context, we extend the PHOG descriptor, which represents spatial local shape, to spatio-temporal domain so as to capture the changes in local shape of facial sub-regions in the temporal dimension to give 3D facial component sub-regions of forehead, mouth, eyebrow and nose. We refer this descriptor as PHOG\_three orthogonal planes (PHOG\_TOP). By combining PHOG\_TOP and dense optical flow of the facial region, we exploit the fusion of discriminant features for classifying and thus recognising facial expressions.

The main contributions of Framework 2 are as follows: (a) a framework that integrates the dynamic information extracted from variation in facial shape and movement of facial landmark; (b) PHOG\_TOP 3D facial features; (c) a means of using weighted PHOG\_TOP with dense optical flow descriptor; and (d) an analysis on the contribution of different facial subregions using Framework 2.

## 4.2 Proposed framework

Framework 2 consists of three modules: pre-processing, feature extraction and classification as shown in Figure 4.1. The pre-processing includes facial landmark detection (face acquisition) and face alignment, where face alignment is applied to reduce the effect of variation in head pose and scene illumination to give a better recognition performance.

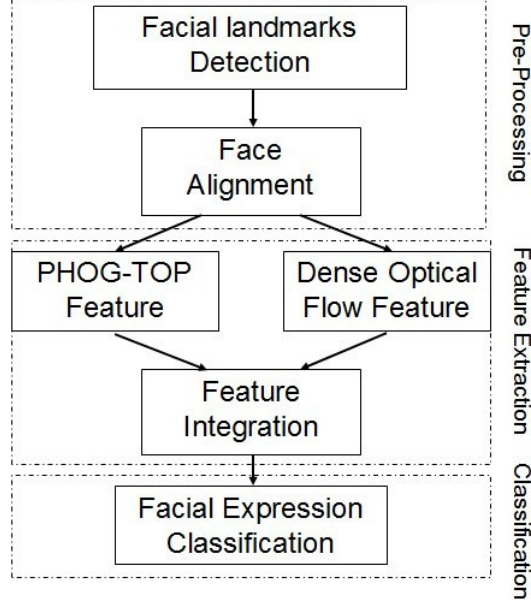


Figure 4.1: Framework 2 for facial expression recognition.

The Local Evidence Aggregated Regression (LEAR) [180] is employed to detect facial landmarks over every image frame of a video sequence, and the locations of the detected eyes are then used to align any in-plane rotation, where the angle of the two eyes in each frame are rotated at their centre to line up to the horizontal axis. The two eyes and nose tip are used to scale and crop each image frame into a  $160 \times 240$  rectangular region of interest containing the central face region. In the cropped image, the  $x$  coordinate of the centre of the two eyes are the centre in the horizontal direction, while the  $y$  coordinate of the nose tip locates the lower third in the vertical direction.

The feature extraction phase includes the generation of PHOG\_TOP and  $v_{\text{global}}$ . The details are presented in Section 4.3.

After the integrated descriptor is obtained, two popular classifiers (SVM and K-Nearest Neighbour (KNN) Classifier) are introduced in Framework 2 for classification.

KNN classifier is one of widely used instance-based classification methods due to its simplicity. KNN aims to compute the similarity between all subjects in the training samples and a test sample, and utilises a majority vote scheme to determine the class of the current test sample. One advantage of KNN is that it is suitable for multi-class classification task since its classification is based on a small neighbourhood of similar object. The most commonly used similarity measure

## 4.2. PROPOSED FRAMEWORK

---

of KNN classifier is the Euclidian distance metric. Given two feature vector  $\mu = (\mu_1, \mu_2, \dots, \mu_m)$  and  $v = (v_1, v_2, \dots, v_m)$ , the Euclidean distance between these two vectors is

$$d(\mu, v) = \sqrt{\sum_{i=1}^m (\mu_i - v_i)^2}, \quad (4.1)$$

where  $m$  is the size of vector.

SVM has been widely applied to facial expression recognition due to its following properties: (1) its ability to work with high-dimensional data; and (2) a high generalisation performance without the need of a priori knowledge, even when the dimension of the input space is very large. The linear, polynomial, and RBF kernels have been shown in [89] to achieve good performance in facial expression recognition. Thus, to achieve an effective classification of facial expressions, the classical SVM classifier with the RBF kernel is used in Framework 2, where grid-search and 10-fold cross-validation [147] are used to estimate the kernel parameter. The parameter that achieves the best cross-validation accuracy is selected.

SVM is a binary classifier, however, the classification of facial expressions is a multi-class classification problem. Thus, the binary SVM classifiers need to be combined for recognition of multiple classes [147]. Two strategies are commonly used: one-versus-one and one-versus-all. In the one-versus-all strategy, the classifier with the highest output assigns the class. In the one-versus-one case, every classifier assigns test data to one of the two classes, the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the classification. One-versus-one strategy is used in Framework 2 due to the simplicity of its implementation and its robustness in classification as follows. Suppose  $C$  classes are classified using binary classifiers,  $C(C-1)/2$  binary classifiers are built from all pairs of distinct classes. Sometimes, more than one expressions (i.e., classes) obtain the most number of votes. In this case the features extracted from image sequences of one of the two classes, say class  $m$ , are averaged to obtain a template representing that class, i.e.,

$$T_m = \frac{\sum_i^M f_{i,m}}{M}, \quad (4.2)$$

where  $f_{i,m}$  denotes the  $i$ th feature belonging to class  $m$ ,  $M$  is the number of features which belongs to class  $m$ . During the classification, a simple nearest neighbour classifier is used and the feature belonging to the test data  $h_1$  is classified as the nearest class template, i.e.,

$$T_m : d(h_1, T_m) < d(h_1, T_n), \quad (4.3)$$

### 4.3. DYNAMIC FEATURES EXTRACTION

---

where  $d(.,.)$  is Euclidean distance function, and  $m$  and  $n$  are the indices of the two classes with the most votes.

## 4.3 Dynamic features extraction

Framework 2 uses dynamic features. This takes the form of PHOG\_TOP and dense optical flow, combined to give a robust and accurate recognition of facial expressions.

### 4.3.1 PHOG\_TOP descriptor

PHOG [99], originally designed for object classification, contains the local shape information of an image and spatial layout of this shape. This descriptor was inspired by HOG [98] and the image pyramid representation [96]. In essence, PHOG is a descriptor based on edge information. More specifically, edge contours of an image are extracted at different pyramid resolution level, and occurrences of gradient orientation of edges are counted to construct a gradient histogram. The PHOG descriptor is obtained by concatenating the histograms from selected pyramid levels. An example of a PHOG descriptor of a face is shown in Figure 4.2.

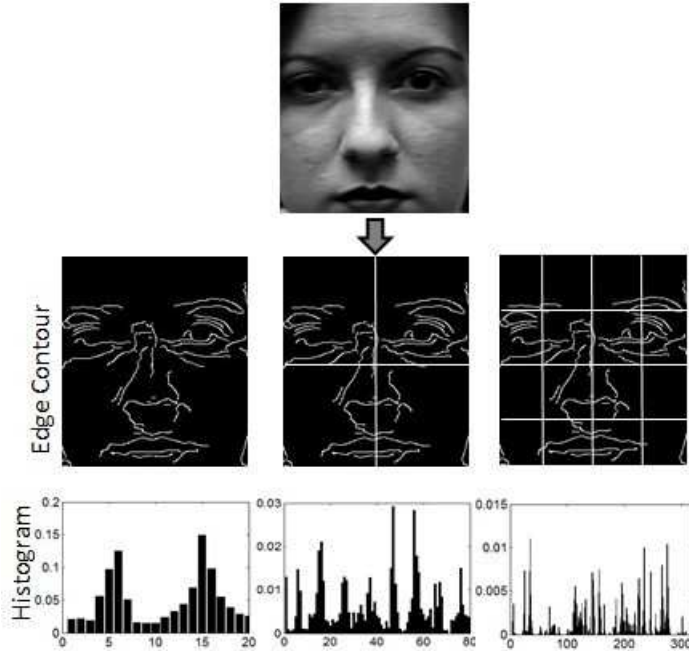


Figure 4.2: PHOG descriptor of a face.

Facial expression is usually performed dynamically, thus its dynamic information is essential for its recognition. Motivated by the temporal extension of

### 4.3. DYNAMIC FEATURES EXTRACTION

LBP [88], this chapter proposes a spatial-temporal descriptor by concatenating the PHOG of three orthogonal planes XY, XT and YT to give PHOG\_TOP, taking into account the co-occurrence statistics in these three planes. The XY plane is used to extract the local spatial information, and the XT and YT planes are used to extract temporal information. Framework 2 considers a video sequence as a stack of XY slices in the temporal dimension, and similarly for XT and YT slices but in the Y and X dimensions, respectively. The spatio-temporal PHOG over each slice in three orthogonal axes (i.e., XY\_PHOG, XT\_PHOG, and YT\_PHOG) are separately obtained and then combined. Take XY\_PHOG for instance, the proposed framework computes PHOG descriptor in every single image from a video sequence, i.e., along the temporal axis. First, the Canny edge detector is employed to capture the edge information. The image region is divided into a set of spatial grid by repeatedly doubling the number of divisions along each axis. Thus, the grid at resolution level  $l$  has  $2^l$  cells along each dimension.

The orientation of gradient for each grid at each resolution level is computed using a Sobel mask without Gaussian smoothing, as the smoothing decreases the performance of classification [99]. The histogram of edge orientations within an image sub-region is quantized into  $K$  bins, where  $K$  is set to 20 as in [99]. In order to reflect the contribution of each edge, a weight proportional to its magnitude is added. Each bin in the histogram represents the occurrences of edges that have orientations within a certain angular range. Framework 2 uses the range  $[0, 360]$  to take into account all orientations. The PHOG descriptor for a slice (image) is obtained by concatenating all the vectors at each pyramid resolution. The final PHOG\_XY is obtained by averaging the PHOG features over all slices in the temporal dimension. The creation of the PHOG descriptor is illustrated in Figure 4.3.

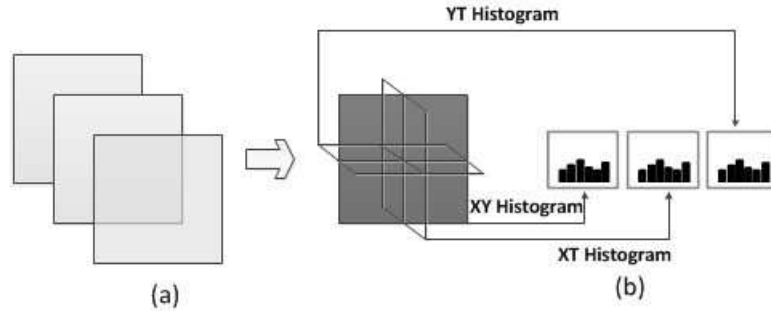


Figure 4.3: Three planes in spatio-temporal domain for extracting TOP features, and the histogram concatenated from three planes: (a) original image, (b) the x-y, y-t, and x-t planes, and the concatenation of resulting histograms into a single feature set.

### 4.3. DYNAMIC FEATURES EXTRACTION

---

The PHOG is normalised to sum to unity. Consequently, level 0 is represented by a  $K$  vector corresponding to  $K$  bins of the histogram, level 1 by a  $4K$ -vector, and the normalised PHOG\_XY descriptor of the entire sequence is the vector

$$\text{PHOG\_XY}_{\text{Seq}} = \frac{\sum_k \text{PHOG\_XY}_k}{K} \quad (4.4)$$

with dimensionality

$$\text{Dim}_{xy} = K \sum_{l \in L} 4^l, \quad (4.5)$$

where  $L$  denotes the number of levels, which is set to 2 to prevent over fitting of the edge contours over the grid. PHOG\_TOP is a concatenation of the descriptors of three planes (PHOG\_XY<sub>Seq</sub>, PHOG\_XT<sub>Seq</sub>, and PHOG\_YT<sub>Seq</sub>, resulting in

$$\text{PHOG\_TOP} = \{\text{PHOG\_YT}_{\text{Seq}}, \text{PHOG\_XT}_{\text{Seq}}, \text{PHOG\_XY}_{\text{Seq}}\} \quad (4.6)$$

of dimensionality  $\text{Dim}_{xy} + \text{Dim}_{xt} + \text{Dim}_{yt}$ .

In [181], the authors segmented an image into a number of sub-regions, which showed a better performance than using the whole image. Following the strategy of [181], Framework 2 segments an image sequence into a number of 3D facial component sub-regions (forehead, mouth, eyebrow and nose) to enhance the spatial information. However, unlike in [181] the framework does not subdivide the video sequence in the temporal dimension. This is because the length of the video sequences (as used in the experiments for this chapter which starts with the neutral expression and ends with the peak phase, i.e., with the most significant motion) extracted from Extended CK+ dataset [94] and MMI dataset [149] are relatively short. Figure 4.4 shows the four facial sub-regions and the 3D sub-region of mouth in a video sequence. PHOG\_TOP is applied to the entire face video sequence and four different sub-regions.

#### 4.3.2 Dense optical flow descriptor

Optical flow captures the dynamic information in a video sequence. To enhance the dynamic information, optical flow is used to track facial points in a video sequence and compute the displacement which represents the movement of corresponding points between two consecutive frames. Instead of tracking some facial fiducial points (landmarks), the framework tracks dense facial points uniformly distributed on a grid placed on the central facial region. The merit of the dense optical flow is that the points on the grid are tracked as one entity. As a result, the global displacement of these considered points is obtained, which is used to generate the



Figure 4.4: (Left) four facial sub-regions, and (right) face video sequence with 3D mouth sub-region.

feature vector.

The Lucas-Kanade optical flow algorithm [177] has been used to estimate the displacement of facial feature points. It is a gradient-based method for motion estimation, and approximates the motion between two consecutive frames. Given two consecutive frames  $I_{t-1}$  and  $I_t$ , for a point  $p = (x, y)^T$  in  $I_{t-1}$ , if the optical flow is  $d = (u, v)^T$  then the corresponding point in  $I_t$  is  $p + d$ , where  $T$  is the transpose operator. The algorithm find the  $d$  which minimises the match error between the local appearances of two corresponding points. A cost function  $e(d)$  is defined for the local area  $R(p)$  [177], i.e.,

$$e(d) = \sum_{x \in R(p)} w(x) (I_t(x + d) - I_{t-1}(x))^2, \quad (4.7)$$

where  $w(x)$  is a weights window, which assigns larger weight to pixels that are closer to the central pixel as these pixels give more importance than others. Optimising (5.5) gives the solution [177]

$$d = G^{-1}H \quad (4.8)$$

where

$$G = \sum_{x \in N(p)} w(x) \nabla I_t (\nabla I_t)^T \quad (4.9)$$

$$H = \sum_{x \in N(p)} w(x) \nabla I_t \Delta I \quad (4.10)$$

$$\Delta I = I_{t-1} - I_t \quad (4.11)$$

$$\nabla I_t = \frac{dI_t}{dx}. \quad (4.12)$$

### 4.3. DYNAMIC FEATURES EXTRACTION

---

The efficiency of computing the dense optical flow depends on the grid size. Since a grid with small cells leads to high computation, we use a grid of  $20 \times 15$  placed on the central face region, and its vertices are the facial points to be tracked. The displacement vector of two consecutive frames is then computed on these 300 points, which reduces the computation significantly. The right image of Figure 3.6 shows the trajectories of tracked points from the previous frame of a neutral facial expression.

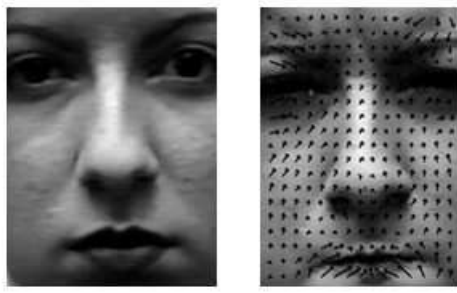


Figure 4.5: (Left) a neutral image, and (right) with the dense optical flow (denoted by needles) superimposed. The magnitude and direction of the flow are respectively indicated by its length and the direction of its arrow.

The corresponding inter-frame displacement vectors are added to obtain the global displacement vectors corresponding to each point during an expression (i.e., from a neutral face to the peak phase of the expression). More formally, let  $(p_1, \dots, p_S)$  be the respective spatial image positions of a facial feature point (i.e., one of the grid vertices)  $p$  at frames  $1, \dots, S$ , where  $S$  is the number of frames in the video sequence. The global displacement vector of point  $p$  is

$$v_p = \sum_{i=1}^{S-1} v_{pi} = \sum_{i=1}^{S-1} (p_{i+1} - p_i), \quad (4.13)$$

where

$$\|v_p\| = \sqrt{p_x^2 + p_y^2}, \quad \theta_p = \tan^{-1}\left(\frac{p_y}{p_x}\right), \quad (4.14)$$

and  $p_x$  and  $p_y$  are respectively the  $x$  and  $y$  components of  $v_p$ .

The normalisation of the dense optical flow is slightly different from that of PHOG\_TOP, where Framework 2 does not detect the landmarks over the entire video sequence. Instead, the framework only locates the landmarks of the first frame of each sequence, and introduces a reference vector for normalisation. More specifically, the two inner eye corner points are used to define a reference vector  $\bar{n}v$



### 4.3. DYNAMIC FEATURES EXTRACTION

---

and its angle  $\theta$  with the horizontal axis  $\bar{w} = (1, 0)^T$ , i.e.,

$$\|\bar{nv}\| = \sqrt{nv_x^2 + nv_y^2} \quad (4.15)$$

$$\theta = \cos^{-1}\left(\frac{\bar{nv} \cdot \bar{w}}{\|\bar{nv}\| \|\bar{w}\|}\right). \quad (4.16)$$

These are used to normalise the global displacement vectors and to correct the orientation angles of these vectors. Finally, the global displacement of any feature point is

$$d_{\text{global}} = \frac{\|v_p\|}{\|\bar{nv}\|}. \quad (4.17)$$

Its normalised angle is computed by adding it to  $\theta_{\bar{p}}$  (i.e., the orientation of vector  $\bar{p}$  is the positive or negative value of  $\theta$ ) to give

$$\theta_{\text{global}} = \theta + \theta_{\bar{p}}. \quad (4.18)$$

Two normalised features are thus obtained for each of the tracked feature points.

A global feature vector  $v_{\text{global}}$  of 600 components, i.e., the dense optical flow, is then created for each video sequence, containing the pair of displacement features of the points, i.e.,

$$v_{\text{global}} = [\|p_1\|, \theta_{p_1}, \|p_2\|, \theta_{p_2}, \dots, \|p_R\|, \theta_{p_R}] , \quad (4.19)$$

where  $\|p_i\|$  and  $\theta_{p_i}$  respectively denote the normalised modulus and the normalised angle of the vector  $p_i$  of accumulated displacement of a given feature point, and  $R$  denotes the number of tracked points.

#### 4.3.3 Integration of descriptors

The integration of the normalised descriptors of a video sequence is

$$\begin{aligned} f_{\text{Int}} &= \{f_{\text{motion}}, f_{\text{spatial}}\} \\ &= \{\omega_1 \cdot \text{PHOG\_YT}_{\text{Seq}}, \omega_2 \cdot \text{PHOG\_XT}_{\text{Seq}}, \\ &\quad \omega_3 \cdot v_{\text{global}}, \omega_4 \cdot \text{PHOG\_XY}_{\text{Seq}}\}, \end{aligned} \quad (4.20)$$

where  $\omega_i$  denote the weights for the different descriptors and represent the contribution of the descriptors to the integrated descriptor. Note that if every weight is set to 1 then the integration is simply the concatenation of all descriptors. The analysis of the selection of weights is presented in Section 5.4. The set of the integrated feature vectors, corresponding to all considered video sequences, is divided into two

disjoint sets: training and test.

## 4.4 Experiments

### 4.4.1 Facial expression datasets

The Extended CK+ dataset (CK+) [94] is the most widely used dataset for evaluating facial expression recognition methods, and is publicly available. This dataset contains 593 image sequences of seven basic facial expressions (namely Anger, Contempt, Disgust, Fear, Happiness, Sadness and Surprise). These expressions were performed by 120 subjects. The age of the participants ranges from 18 to 30 years, 65% of them are female, 81% are Euro-American, 13% are Afro-American, and 6% of other racial groups. Each frame of the image sequences is  $640 \times 480$  or  $640 \times 490$  pixels with an 8-bit grey scale. The video sequences vary in duration (i.e., 10 to 60 frames) and incorporate the onset (i.e., the neutral frame) to peak phase of the facial expression. 327 image sequences of seven expressions were used, where the expression neutral with Contempt. The top row of Figure 4.6 shows sample images of a subject expressing six expressions. Table 4.1 shows the occurrences of the various expression classes in CK+ dataset.

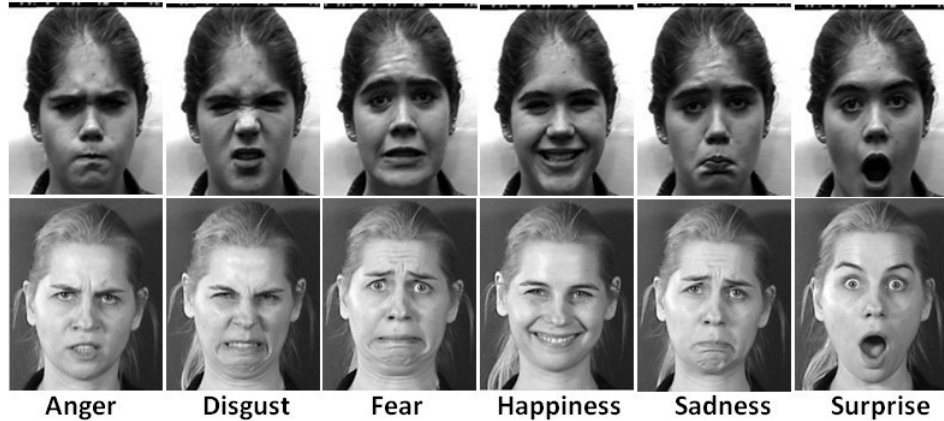


Figure 4.6: Example of expressions from CK+ dataset (top row) and MMI dataset (bottom row). Each image is the frame with the most expressive face in a video sequence.

The MMI dataset [149] is another well-known publicly available dataset, which comprises video sequences including both posed and spontaneous expressions. These expressions were performed by 19 subjects (44% female), with age ranging from 19 to 62, and of European, Asian and South American ethnicity. The subjects performed 79 expressions including the six basic facial expressions with neutral

#### 4.4. EXPERIMENTS

---

Table 4.1: Number of image sequences (subjects) for each expression in the CK+ dataset and MMI dataset.

Expression	CK+	MMI
Anger	45	32
Disgust	59	28
Fear	25	28
Happiness	69	42
Sadness	28	32
Surprise	83	41
Contempt	18	0
Total	327	203

frame at the start of each sequence. Every video frame is at  $720 \times 576$  spatial resolution. The original frames were converted into 8-bit greyscale images for the experiments for this thesis, and the sub-sequence from the neutral frame to the peak phase extracted. In the experiments, 203 image sequences labelled as one of the six basic facial expressions were selected from the MMI dataset. The bottom row of Figure 4.6 shows sample images of the six basic expressions.

Since the MMI dataset was generated in a different way to the CK+ dataset, and may contain larger pose variation, Framework 2 uses slightly different pre-processing to align the data. For dense optical flow, the nose tip is used to detect and subtract the effect of head motion. For PHOG\_TOP, since the sequences selected for the experiments are frontal view with no out-of-plane head pose which can cause self-occlusions of the eyes region (which are used for the alignment process), the framework applies the same pre-processing method as for CK+ dataset.

##### 4.4.2 Experimental results

Three sets of experiments are performed on the CK+ dataset (with smaller head motion than in MMI dataset) to evaluate the performance of Framework 2 and to compare its performance with two state of the art facial expression recognition methods. Leave-sequence-out cross-validation scheme is employed as follows. One video sequence is selected for testing and the remaining video sequences are used for training, which guarantees that a sequence selected for testing is not in the training set and consequently sequence independence is realised in our experiments.

The first set of experiments investigates whether spatio-temporal PHOG performs better than using only one histogram from either XY, YT or XT plane. Cross-validation was applied 327 times on the selected 327 video sequences. The recog-

#### 4.4. EXPERIMENTS

---

recognition rate of classifying all expressions using PHOG from each individual plane, i.e., PHOG\_XY<sub>Seq</sub>, PHOG\_XT<sub>Seq</sub> and PHOG\_YT<sub>Seq</sub>, and their combination, i.e., PHOG\_TOP, are summarized in Figure 4.7.

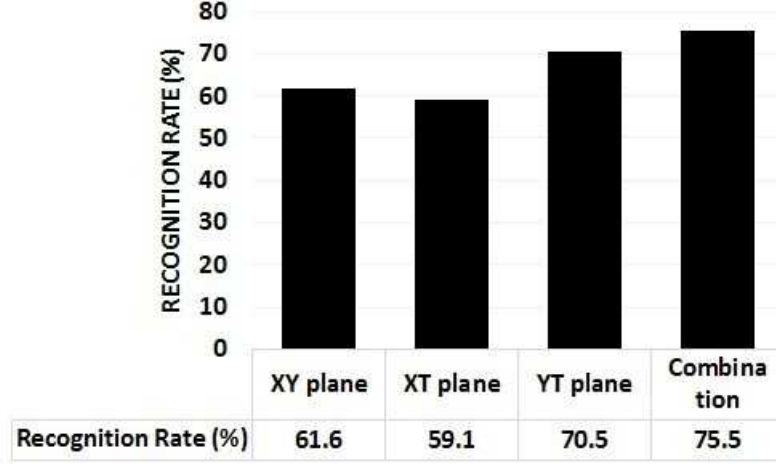


Figure 4.7: Recognition rates of all expressions PHOG from either XY, XT, YT or their combination, i.e., PHOG\_TOP.

Figure 4.7 shows that the recognition rate using the non-weighted (i.e., with  $\omega_i=1$  for  $i=1, 2, 3, 4$ ) combination of features is the highest. Also, using the features from YT plane gives the better performance than from the other two planes, and the features which represent the variation in shape horizontally from the XT plane give the lowest rate. This means: 1) the dynamic motion features (i.e., feature from the YT plane) plays more important role than spatial features in recognizing facial expressions; and 2) the vertical variations in shape (e.g, the opening of mouth) are more significant than horizontal variations.

Framework 2 combines the features extracted from the three planes on the assumption that all components have equal contribution in the recognition. However, from the aforementioned analysis, not all features are equally important, and some of them contain more useful information than others. A weighting strategy is thus introduced to improve the performance in recognition.

In order to determine the appropriate weights, the recognition rates achieved by using features from the three planes separately are analysed so that the features that give better recognition are identified and are allocated bigger weight, i.e.,

$$f_{weighted} = \omega_i f_i \quad (4.21)$$

where  $\omega_i$  and  $f_i$  are the weight of the  $i_{th}$  plane and PHOG extracted from the  $i_{th}$

#### 4.4. EXPERIMENTS

plane, respectively.

The weighting strategy is as follows. First, given the three recognition rates achieved using features from the three planes separately, the framework obtains  $R = [R_1, R_2, R_3]$  from the output of the SVM classifier, where the lower the rate the smaller contribution the feature has. The weight vector

$$\omega_i = \frac{\|R_i\|}{\|R\|}, \quad (4.22)$$

where  $\|\cdot\|$  is  $L_2$  norm,  $R_i$  denotes the average recognition rate using features from the  $i_{th}$  plane, and  $R = [R_1, R_2, R_3]$ .

The recognition rates achieved using combination of non-weighted and weighted features are shown in Figure 4.8. The use of the combination of weighted features resulted in better performance for sadness and surprise, and similar performance for disgust and contempt. This is because the variations in shape in sadness and surprise are more significant.

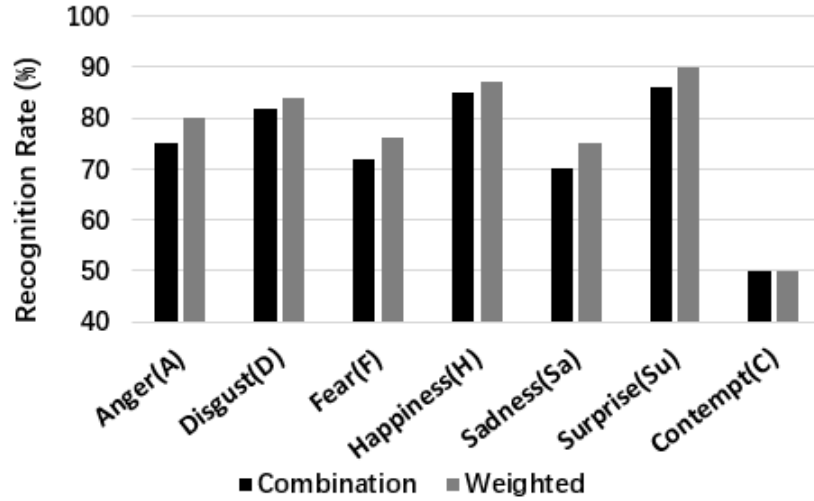


Figure 4.8: Recognition rates of all expressions in using combination of non-weighted and weighted PHOG\_TOP.

The second set of experiments evaluates the discriminant power of the different facial regions (i.e., how discriminative different facial regions are) and determines how much the PHOG\_TOP extracted from different facial region contribute to the six expressions and Contempt. Four facial sub-regions (namely eyebrows, forehead, nose and mouth) are extracted and are used to compute the PHOG\_TOP. The concatenation of these facial sub-regions are also taken into account. Table 4.2 shows the classification of PHOG\_TOP from the whole facial region achieves good recog-

#### 4.4. EXPERIMENTS

---

nition rate (i.e.,  $\geq 85\%$ ) for Disgust, Happiness and Surprise, but less for Fear and Sadness. This is due: (a) the number of samples for Fear and Sadness are relatively small; (b) Happiness and Surprise were performed with much more distinguishable movement of facial landmarks; and (c) Fear and Sadness were poorly performed that it is difficult even for human to distinguish them. Figure 4.9 shows the average classification of six expressions using PHOG\_TOP for the considered sub-regions and their concatenation. The results show that among the sub-regions the mouth provides the best recognition rate. This is due to the more-differentiated movement of the mouth muscles (and consequently the corresponding facial points) for each facial expression. Figure 4.9 also shows the concatenation of the four sub-regions provides better recognition than using the whole face. This is because the local shape information in both spatial and temporal domains is enhanced, while the other information is removed.

Table 4.2: Multiclass SVM results of PHOG\_TOP from the whole face on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>	<b>C</b>
<b>A</b>	36	1	3	0	3	0	2
<b>D</b>	5	51	0	0	0	2	1
<b>F</b>	2	0	19	0	3	1	0
<b>H</b>	1	1	2	62	0	1	2
<b>S</b>	1	0	3	1	21	1	1
<b>Su</b>	1	0	2	2	1	76	1
<b>C</b>	3	3	0	2	1	0	9

The third set of experiments evaluates the effectiveness of combining spatial shape information with dynamic motion features. Table 4.3, Table 4.4 and Table 4.5 respectively show the results obtained in using dense optical flow, PHOG\_TOP and Framework 2 with multi-class SVM classifier, including the overall classification performance of multi-class SVM, where the dark grey shade indicates the correct recognition result of each emotion while the lighter grey shade indicates the recognition result of each emotion misclassified with the maximal error. The three tables show that the combination of PHOG\_TOP and dense optical flow feature is more accurate than using individual features separately in recognising facial expressions.

#### 4.4. EXPERIMENTS

---

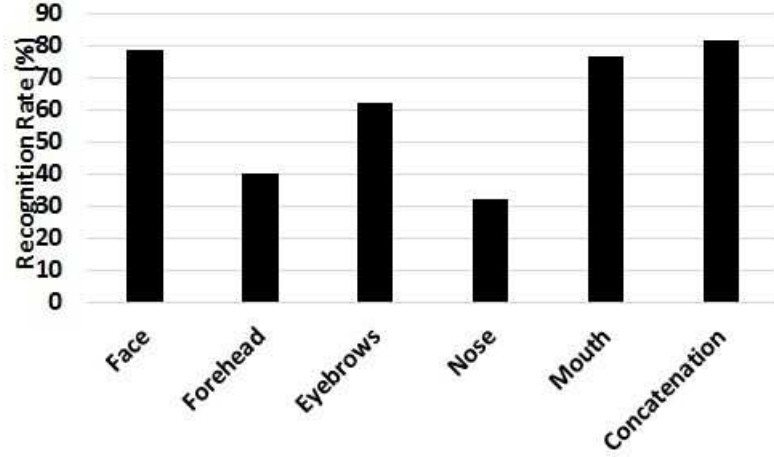


Figure 4.9: Discriminant power of of facial sub-regions in recognising six expressions using PHOG\_TOP.

Table 4.3: Multiclass SVM results in using Dense flow optical flow on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	A	D	F	H	Sa	Su	C	%
A	33	4	0	3	4	0	1	73.3
D	4	50	1	1	0	0	3	84.8
F	3	1	9	5	2	3	2	36.0
H	1	0	2	63	1	0	2	91.3
Sa	4	0	2	2	14	3	3	50.0
Su	1	1	0	1	7	72	1	86.8
C	6	0	2	1	1	0	8	44.4

#### 4.4. EXPERIMENTS

---

Table 4.4: Multiclass SVM results in using PHOG\_TOP on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>	<b>C</b>	<b>%</b>
<b>A</b>	38	1	3	0	2	0	1	84.4
<b>D</b>	4	53	0	0	0	1	1	89.8
<b>F</b>	1	0	21	0	3	0	0	84.0
<b>H</b>	0	1	2	64	0	0	2	92.8
<b>Sa</b>	2	0	3	0	21	1	1	75.0
<b>Su</b>	0	0	1	2	1	79	0	95.2
<b>C</b>	4	2	0	2	1	0	9	50.0

Table 4.5: Multiclass SVM results in using Framework 2 on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>	<b>C</b>	<b>%</b>
<b>A</b>	40	0	1	0	3	0	1	88.9
<b>D</b>	2	55	1	0	0	0	1	93.2
<b>F</b>	0	1	20	1	3	0	0	80.0
<b>H</b>	1	0	2	65	0	1	0	94.2
<b>Sa</b>	2	0	3	0	22	0	1	78.5
<b>Su</b>	0	0	0	1	3	79	0	95.2
<b>C</b>	3	1	1	2	1	0	10	55.6



#### 4.4. EXPERIMENTS

---

It is difficult to make a quantitative comparison between the state of the art facial expression recognition methods due to the different pre-processing and experiment strategies involved. Since the methods of Eskin and Benli [182] and Lucey et al. [94] were evaluated on the CK+ dataset using leave-one-out cross-strategy, it is possible to compare Framework 2 with these methods as shown in Table 4.6. Table 4.6 shows that Framework 2 achieves an average recognition rate for all seven facial expressions of 83.70%, which outperforms the dynamic method of Eskin and Benli [182] and the static method of Lucey et al. [94]. The confusion matrices in Table 4.7, Table 4.8 and Table 4.9 respectively show the recognition rate in using the proposed features of dense optical flow, PHOG\_TOP and the combined optical flow and PHOG\_TOP (i.e., Framework 2), where the number in a parentheses denotes the difference in recognition rate in using the proposed features and those used in Lucey et al [94] (i.e., shape, appearance and combined shape and appearance, respectively), where a positive number indicates the proposed features performs better and a negative number indicate worse. Table 4.7 and Table 4.8 respectively show the use of dense optical flow and PHOG\_TOP achieves significantly better performance than the use of shape and appearance in five of the seven facial expressions. Table 4.9 shows the combined use of dense optical flow and PHOP\_TOP is significantly better in three expressions, slightly worse in two expressions and significantly worse in one expression. However the average recognition performance of Framework 2 is slightly better as shown in Table 4.6. It can also be concluded that the combination of spatial local shape information and dynamic features improves the recognition.

Table 4.6: Comparative evaluation of Framework 2 with 2 methods using leave-subject-out cross-validation.

Study	Methodology	Recognition Rate
Eskin and Benli [182]	SVM	<b>76.8</b>
	Adaboost	<b>76.3</b>
Lucey et al. [94]	SVM(shape)	<b>50.4</b>
	SVM(Appearance)	<b>66.7</b>
	SVM(Combined)	<b>83.3</b>
Framework 2	SVM	<b>83.7</b>

#### 4.4. EXPERIMENTS

---

Table 4.7: Confusion matrix in using dense flow optical flow on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using dense optical flow and in using shape.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>
<b>A</b>	73.3(38.3)	8.9	0.0	6.7	8.9	0.0
<b>D</b>	6.8	84.8(16.4)	1.7	1.7	0.0	0.0
<b>F</b>	12.0	4.0	36.0(14.3)	20.0	8.0	12.0
<b>H</b>	1.4	0.0	2.9	91.3(-7.1)	1.4	0.0
<b>Sa</b>	14.3	0.0	7.1	7.1	50.0(46.0)	10.7
<b>Su</b>	1.2	1.2	0.0	1.2	8.4	72(-28.0)

Table 4.8: Confusion matrix in using PHOG\_TOP on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using PHOG\_TOP and in using appearance (of Lucey et al.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>
<b>A</b>	84.4(14.4)	2.2	6.7	0.0	4.4	0.0
<b>D</b>	6.8	53(41.7)	0.0	0.0	0.0	1.7
<b>F</b>	4.0	0.0	84.0(62.3)	0.0	12.0	0.0
<b>H</b>	0.0	1.4	2.9	64.0(-36.0)	0.0	0.0
<b>Sa</b>	7.1	0.0	10.7	0	75.0(15.0)	3.6
<b>Su</b>	0	0	1.2	2.4	1.2	95.2(-0.8)

#### 4.4. EXPERIMENTS

Table 4.9: Confusion matrix in using Framework 2 on classification of six facial expressions and Contempt of the CK+ dataset with leave-sequence-out cross-validation. The value in a parentheses denotes the difference in recognition rate in using the combined features of dense optical flow and PHOG\_TOP and in using the combined features of shape and appearance (of Lucey et al.) Dark grey shade - correct recognition result of each emotion; and lighter grey shade - recognition result of each emotion misclassified with the maximal error.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>Sa</b>	<b>Su</b>
<b>A</b>	88.9(13.9)	<b>0.0</b>	<b>2.2</b>	<b>0.0</b>	<b>6.7</b>	<b>0.0</b>
<b>D</b>	<b>3.4</b>	93.2(-1.5)	<b>1.7</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>
<b>F</b>	<b>0.0</b>	<b>4.0</b>	80.0(14.8)	<b>4.0</b>	<b>12.0</b>	<b>0.0</b>
<b>H</b>	<b>1.4</b>	<b>0.0</b>	<b>2.9</b>	94.2(-5.6)	<b>0.0</b>	<b>1.4</b>
<b>Sa</b>	<b>7.1</b>	<b>0.0</b>	<b>10.7</b>	<b>0.0</b>	78.5(10.5)	<b>0.0</b>
<b>Su</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	<b>1.2</b>	<b>3.6</b>	95.2(-0.8)

The MMI dataset is used to provide quantitative comparisons with the methods developed by Shan et al. [89], Fang et al. [35], AAM [35], and ASM [35]. For the experiments for this thesis, 10-fold cross-validation was performed. The average recognition rates are shown in Table 4.10. The performance of Framework 2 on MMI dataset is worse than that on the CK+ dataset. This is because there are fewer data for training and there are larger changes in head pose in MMI dataset. However, Framework 2 outperforms all the other four methods.

To evaluate the across-dataset performance of Framework 2, the integrated features from CK+ dataset are extracted as training data, and the MMI dataset used for testing. The recognition result (in Table 4.10) shows the generalisation performance across datasets is much lower (58.7% on the MMI database). Thus, it is concluded that the current expression classification trained on a single dataset under controlled environment gives good performance only on that dataset.

The computational complexity is analysed with the view of assessing the potential of Framework 2 for real-time application. The processing time is approximated to be the time needed for preprocessing, extracting the dense optical flow and PHOG\_TOP, and classification per image frame of the video sequence. The processing time (measured using the computer system clock) is estimated using OpenCV 2.4.3 in Microsoft Visual Studio 2010 Express Edition environment on an Intel(R) Core (TM) i7-3770 CPU @ 3.40ghZ with 16GB RAM running on Windows 7 operating system. The average processing time per image is under 350 milliseconds and

Table 4.10: Comparative evaluation of Framework 2 using MMI dataset

<b>Study</b>	<b>Recognition Rate</b>
Shan et al. [89]	<b>54.45</b>
AAM in [35]	<b>62.38</b>
ASM in [35]	<b>64.35</b>
Fang et al. [35]	<b>71.56</b>
Framework 2	<b>74.30</b>
train: CK+ Test: MMI	<b>58.70</b>

520 milliseconds for CK+ and MMI datasets, respectively.

## 4.5 Conclusion

This chapter presents a facial expression recognition framework which uses PHOG\_TOP and dense optical flow. The proposed Framework 2 which comprises pre-processing, feature extraction and multi-class SVM-based feature classification achieves better performance than two state of the art methods on CK+ dataset and four other methods on MMI dataset. The average recognition rate is 83.7% on the CK+ dataset and 74.3% on the MMI dataset. The expressions of Happiness and Surprise are easier to be distinguished than other facial expressions, and the results on these two expressions demonstrate the capability of Framework 2. Nevertheless, the recognition rate of the expressions of Contempt (55.6%) and Sadness (78.5%) on the CK+ dataset are lower because these two expressions are often misclassified as Anger and Fear. A limitation of Framework 2 is its generalisation to other datasets. This is because different datasets are generated under different environments.

Framework 2 is able to classify the expressions of Happiness and Surprise accurately, but encountered some difficulty in recognising the other expressions. This is because the datasets used in this thesis are small and some of the expressions are poorly represented. One way to address this is to either balance the dataset, or acquire more data. Also, Framework 2 only focuses on the change of geometric features (i.e. movement of landmarks and variation of shape) neglecting the changes of texture and appearance information (i.e. wrinkle of forehead and furrow) when showing expression. Thus, combining Framework 2 with the change of appearance information might improve the performance of recognition.

## Chapter 5

# Spatio-temporal Framework Based on Local Zernike Moment and Motion History Image

This chapter presents a spatio-temporal feature based on local Zernike moment in the spatial domain using motion change frequency. A novel dynamic feature comprising motion history image and entropy is also designed. To recognise a facial expression, a weighting strategy based on the latter feature and sub-division of the image frame is applied to the former to enhance the dynamic information of facial expression, and followed by the application of the classical support vector machine. Experiments on the CK+ and MMI datasets using leave-one-out cross validation scheme demonstrate that the integrated framework achieves a better performance than using individual descriptor separately. Compared with six state-of-the-art methods, Framework 3 which is proposed in this chapter demonstrates a superior performance.

### 5.1 Introduction

Shape as a geometric-based representation is crucial for interpreting facial expressions. However, current state-of-the-art methods only focus on a small subset of possible shape representation, e.g., point-based methods that represent a face using the locations of several discrete points. Noting that image moments can describe simple properties of a shape, e.g., its area (or total intensity), its centre and its orientation, Zernike moments (ZMs) have been used to represent a face and facial expressions in [183; 184]. In [185], Quantised Local Zernike Moment (QLZM) is

used to describe the neighbourhood of a face sub-region. QLZM can capture the local shape information in the spatial domain, but neglecting the dynamic information in the spatio-temporal domain. the proposed framework design a effective way to extended this spatial representation into spatio-temporal which can exploit the dynamics, which improves the recognition performance significantly.

Since a facial expression involves a dynamic process, and the dynamics contain information that represents a facial expression more effectively, it is important to capture such dynamic information so as to recognise facial expressions over the entire video sequence. Recently, there has been more effort on modelling the dynamics of a facial expression sequence. However, the modelling is still a challenging problem. Thus, this chapter focuses on analysing the dynamics of facial expression sequences. First, the spatial domain QLZM descriptor is extended into spatio-temporal domain, i.e., Motion Change Frequency based QLZM (QLZM\_MCF), which enables the representation of temporal variation of expressions. Second, optical flow is applied to Motion History Image (MHI) [186], i.e., (optical flow based MHI) MHI\_OF, to represent spatial-temporal dynamic information (i.e., velocity), where the proposed MHI\_OF not only captures the movement of facial parts, but also the speed of movement.

Entropy-based methods extract intensity information of image pixels, and have been applied for face recognition. For example, Cament et al. [136] combined entropy-like weighted Gabor features with the local normalisation of Gabor features. Chai et al. [137] introduced the entropy of a facial region, where a low entropy value means the probabilities of different intensities are different, and a high value means the probabilities are the same. They used the entropy of each of the equal-size blocks of a face image to determine the number of sub-blocks within each block. Inspired by [137], Framework 3 uses entropy in the proposed MHI\_OF as follows. Since the intensity value of each pixel in MHI represents a movement, the high intensity values denoting large movement will result in high entropy value, and vice versa.

The proposed Framework 3 utilises two types of features: a spatio-temporal shape representation, QLZM\_MCF, to enhance the local spatial and dynamic information, and a dynamic appearance representation, MHI\_OF. The framework also introduces an entropy-based method to provide spatial relationship of different parts of a face by computing the entropy value of different sub-regions of a face. The main contributions of this chapter are: (a) QLZM\_MCF; (b) MHI\_OF; (c) an entropy-based method for MHI\_OF to capture the motion information; and (d) a strategy integrating QLZM\_MCF and entropy to enhance spatial information.

## 5.2 Proposed framework

Figure 5.1 outlines the proposed Framework 3 which comprises pre-processing, feature extraction and classification. The pre-processing includes facial landmark detection and face alignment, where face alignment is applied to reduce the effects of variation in head pose and scene illumination. The framework uses local evidence aggregated regression [187] to detect facial landmarks over each frame, where the locations of detected eyes and nose are used for face alignment including scaling and cropping. The aligned face images are the size of  $200 \times 200$ , where the x coordinate of the centre of the two eyes are the centre in the horizontal direction, while the y coordinate of the nose tip locates the lower third in the vertical direction. Since the dimensionality of the features is high, following the feature extraction as in Section 5.3 a dimension reduction technique is applied to obtain a more compact representation. Different classifiers may lead to different recognition performance. Framework 3 uses SVM that has been widely used and shown to be effective in recognising facial expressions.

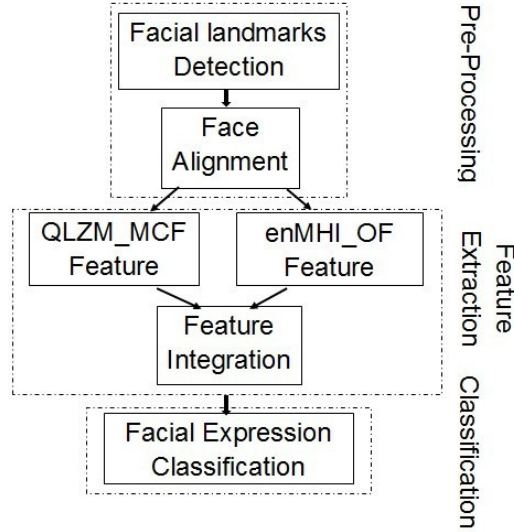


Figure 5.1: Framework 3 for facial expression recognition.

### 5.2.1 Dimensionality reduction using 2D PCA

Principal Component Analysis (PCA) is widely used in facial expression recognition for reducing the dimensionality of feature space. It aims to extract decorrelated features out of possible correlated features using a linear mapping function. Under controlled head-pose and imaging conditions, these features capture the statistical

### 5.3. FEATURE EXTRACTION

---

structure of facial expressions. 2D PCA has been shown to be superior to PCA in terms of more accurate estimation of covariance matrices and reduced computational complexity for feature extraction by operating directly on 2D matrices instead of 1-dimensional vectors [188], i.e., it is not necessary to convert the 2D image into 1D feature prior to feature extraction. Given  $L$  training samples, i.e.,  $G_1, G_2, \dots, G_L$ , the scatter matrix  $S$  is [188]

$$S = \frac{1}{L} \sum_{i=1}^L (G_i - M)^T \times (G_i - M) \quad (5.1)$$

where  $M = (1/L) \sum_{i=1}^L G_i$ . Since there are at most  $L - 1$  eigenvectors of  $S$  with non-zero eigenvalues,  $N$  eigenvectors (where  $N < L - 1$ ) are randomly chosen from the set of  $L - 1$  eigenvectors, i.e.,  $(e_1, e_2, \dots, e_{L-1})$ , with the largest eigenvalues used to construct  $L$  subspaces  $R_{k=1}^L$ . The  $n$ th eigenvector with zero eigenvalue is discarded in order to reduce the dimensionality of the feature space while preserving discriminatory information. Thus, 2D PCA is adopted in this chapter.

## 5.3 Feature extraction

### 5.3.1 Motion History Image

MHI can be considered as a two-component temporal template, a vector-valued image where each component of each pixel is some function of the motion at that pixel location. The MHI  $H_\tau(x, y, t)$  is computed from an update function  $\Psi(x, y, t)$ , i.e.,

$$H_\tau(x, y, t) = \begin{cases} \tau, & \Psi(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t) - \delta), & \text{otherwise} \end{cases} \quad (5.2)$$

where  $(x, y, t)$  is the spatial coordinates  $(x, y)$  of an image pixel at time  $t$  (in terms of image frame number), the duration  $\tau$  determines the temporal extent of the movement in terms of frames, and  $\delta$  is the decay parameter.  $\Psi(x, y, t)$  is defined as

$$\Psi(x, y, t) = \begin{cases} 1, & D(x, y, t) \\ 0, & \text{otherwise} \end{cases} \quad (5.3)$$

where  $D(x, y, t)$  is a binary image comprising pixel intensity differences of frames separated by temporal distance  $\Delta$ , i.e.,

$$D(x, y, t) = |I(x, y, t) - I(x, y, t \pm \Delta)| \quad (5.4)$$



### 5.3. FEATURE EXTRACTION

---

and  $I(x, y, t)$  is the intensity value of pixel with coordinates  $(x, y)$  at the  $t$ th frame of the image sequence. The duration  $\tau$  and the decay parameter  $\delta$  have an impact on the MHI image. If  $\tau$  is smaller than the number of frames, then the prior information of the motion in its MHI will be lost. For example, when  $\tau = 10$  for a sequence with 19 frames, the motion information of the first 9 frame will be lost if the value of  $\delta = 1$ . On the other hand, if the temporal duration is set at very high value compared to the number of frames, then the changes of pixel value in the MHI is less significant. The MHI of a sequence from the Extended CK dataset (CK+) [189] is shown in Figure 5.2.



Figure 5.2: Example of images from sequences (left and middle) and its MHI (right).

#### 5.3.2 Optical Flow Algorithm

Optical flow descriptor can represent the velocity of a set of individual pixels in an image, which capture their dynamic information. Optical flow descriptor is employed in Framework 3 to exploit velocity information.

The Lucas-Kanade method is one of most widely-used method for optical flow computation [190], which solves basic optical flow equation for all pixels in their local neighbourhood by using the least squares criterion. As presented in Section 4.3.2 (presented again here for ease of referral within this chapter), given two consecutive image frames  $I_{t-1}$  and  $I_t$ , for a point  $p = (x, y)^T$  in  $I_{t-1}$ , if the optical flow is  $d = (u, v)^T$  then the corresponding point in  $I_t$  is  $p + d$ , where  $T$  is the transpose operator. The algorithm finds the  $d$  which minimises the match error between the local appearances of two corresponding points. A cost function is defined for the local area  $R(p)$ , i.e., [190]

$$e(d) = \sum_{x \in R(p)} w(x) (I_t(x + d) - I_{t-1}(x))^2, \quad (5.5)$$

where  $w(x)$  is a weights window, which assigns larger weight to pixels that are closer to its central pixel as these pixels give more importance than others.

### 5.3.3 Optical Flow based MHI (MHI\_OF)

Framework 3 computes the optical flow between two consecutive frames and obtains the optical flow image where the intensity of each pixel represents the magnitude of optical flow descriptor. The higher values denote the faster movement of facial points. The framework defines MHI\_OF of a sequence as

$$M(x, y, t) = d(x, y, t) + M(x, y, t - 1) * \bar{\tau} \quad (5.6)$$

where  $\bar{\tau}$  is another decay parameter, and

$$d(x, y, t) = \begin{cases} a * d(x, y, t) + b & d(x, y, t) > T \\ 0 & \text{otherwise.} \end{cases} \quad (5.7)$$

$a$  and  $b$  are scale factors, and  $T$  is a threshold which is used to remove small movements, while retaining large movements of some fiducial points (e.g., eyebrows, lips, etc.). Scale factors are used because the optical flow descriptor is not significantly large for the movements of points in two consecutive frames. In the experiments for this chapter, the original optical flow  $d(x, y, t)$  is magnified by a scale factor  $a$  of 10 with a starting value  $b$  of 20, and the threshold  $T$  is set to 1. A large value of the decay parameter  $\bar{\tau}$  creates a slow decrement of the accumulated motion strength, and the long-term history of the motion is recorded in the resulting MHI\_OF image. A small value of  $\bar{\tau}$  gives an accelerated decrement of motion strength, and only the recent short-term movements are retained in the MHI\_OF image. Figure 5.3 illustrates optical flow based MHI for some facial expressions.



Figure 5.3: Optical flow based MHI for Anger, Happiness and Surprise (from right to left).

#### 5.3.4 Entropy

The entropy of a discrete random variable  $X$  with possible values  $\{x_0, x_1, \dots, x_N\}$  can be defined as [191]

$$E(X) = - \sum_{i=0} p(x_i) \times \log_2(p(x_i)), \quad (5.8)$$

where  $p(\cdot)$  is the probability function. For a grey-level face image, the intensity value of each pixel varies from 0 – 255, and the possibilities of a particular value occurring is random and varies depending on the pattern of different face images. Considering a face image with dimension  $H \times W$  having a total of  $M = H \times W$  pixels, the probability of a particular intensity value  $x_i$  occurring in the image is  $p(x_i) = n_i/M$ , where  $n_i$  is the number of occurrences of  $x_i$  among the  $M$  pixels. In this case, considering  $\sum_i n_i = M$ , the entropy of the image can be expressed as

$$E(X) = \log_2 M - \frac{1}{M} \times \sum_{i=0}^{255} n_i \times \log_2(n_i). \quad (5.9)$$

It is shown in [192] that certain facial regions contain more important information for recognising facial expressions than others. For example the regions of mouth and eyes that produce more changes than those of nose and forehead during an expression have more contribution towards the recognition. Also, as can be seen from the leftmost and middle columns of Figure 5.4, different facial regions in MHI have different intensity levels due to the distance and speed of movements during an expression. Thus, introducing a weight function which allocates different weights to different facial regions will improve recognition. Instead of setting weights empirically based on the observation, Framework 3 utilises entropy to determine the weights as it is expected that the entropy at different facial regions will differ significantly due to pixel intensity variation at these regions.

The size of the training samples in practice is often not large enough to cover all the possible values of pixels in MHI. To address this sparse problem, Framework 3 divides the possible 256 intensity levels into several sections to form intensity divisions. For a 2-dimensional (2D) matrix  $X = (x_{ij})_{H \times W}$ , let  $\chi = \{t_1, t_2, \dots, t_K\}$  be the sorted set of all possible  $K$  intensity values that exist in  $X$  where  $t_1 < t_2 < t_3 \dots < t_K$  and  $K$  is the number of the distinct intensity values. The process of

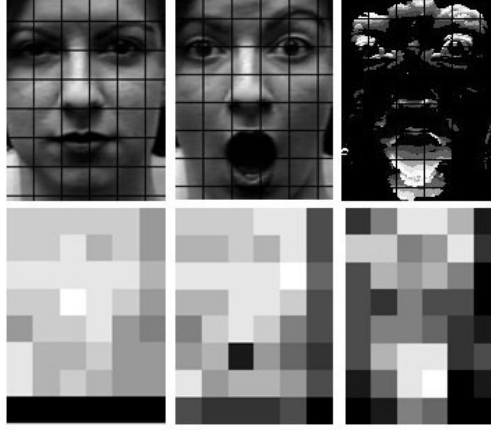


Figure 5.4: Example image entropies: (left column) neutral image and its entropy; (middle column) surprise image and its entropy; and (right column) MHI of surprise image and its entropy. Lighter shades denote larger entropy values.

division is

$$x_{ij} = \begin{cases} x_{t_1}, & t_1 \leq x_{ij} \leq t_2 \\ x_{t_2}, & t_2 \leq x_{ij} \leq t_3 \\ x_{t_3}, & t_3 \leq x_{ij} \leq t_4, \\ & \cdot \\ & \cdot \\ & \cdot \\ x_{t_k}, & t_{K-1} \leq x_{ij} \leq t_K. \end{cases} \quad (5.10)$$

To compute the weight function, Framework 3 divides the MHIs with size of  $H \times W$  into several non-overlapping sub-regions. The 2D spatial histogram of the intensity values  $x_{t_k}$  on each sub-region of  $X$  is

$$h_k = \{h_k(p, q) | 1 \leq p \leq P, 1 \leq q \leq Q\}, \quad (5.11)$$

where  $p, q \in \mathbb{Z}^+$ ,  $P \times Q$  is the size of sub-regions, and  $h_k(p, q) \in [0, \mathbb{Z}^+]$  is the number of occurrences of the intensity values  $x_{t_k}$  in the spatial grid located on the image sub-region of  $[(p-1)\frac{H}{P}, p\frac{H}{P}] \times [(q-1)\frac{W}{Q}, q\frac{W}{Q}]$ . In forming 2D spatial histogram  $h_k$  of intensity values  $x_{t_k}$ , the aspect ratio of the original image is maintained on spatial grids. In this way, spatial characteristics of pixels are retained when forming the 2D spatial histogram.

The entropy value on each sub-region of the 2D spatial histogram is computed

### 5.3. FEATURE EXTRACTION

---

for intensity values  $x_{t_k}$  using

$$S(p, q) = - \sum_{k=1}^K p(h_k(p, q)) \log_2 p(h_k(p, q)), \quad (5.12)$$

where  $p(h_k(q, p))$  is the possibility of particular intensity value  $x_{t_k}$  in the spatial grid located on the image sub-region of  $[(p-1)\frac{H}{M}, p\frac{H}{M}] \times [(q-1)\frac{W}{N}, q\frac{W}{N}]$ .

The normalisation process is implemented using

$$\omega(p, q) = (s(p, q) - s_{\min}) / (s_{\max} - s_{\min}) \quad (5.13)$$

to convert the range of weights into 0-1, where  $s_{\min}$  and  $s_{\max}$  are respectively the maximum value and the minimum of the entropy values over all sub-regions. The computed weights of each subregion on MHI\_OF are as the final weight features

$$\text{enMHI\_OF} = \{\omega(1, 1), \omega(1, 2), \dots, \omega(1, q), \dots, \omega(p, q)\}. \quad (5.14)$$

The MHI\_OF using entropy representation is shown in the rightmost column of Figure 5.4.

#### 5.3.5 Local Zernike Moment

ZMs of an image is computed by decomposing the image onto a set of complex orthogonal basis on the unit disc  $x^2 + y^2 \leq 1$  called Zernike polynomials. The Zernike polynomials are defined as [185]

$$V_{nm}(\rho, \theta) = V_{mn}(\rho \cos \theta, \rho \sin \theta) = R_{nm}(\rho) e^{jm\theta}, \quad (5.15)$$

where  $n$  is the order of the polynomial and  $m$  is the number of iterations such that  $|m| < n$  and  $n - |m|$  is even.  $R_{mn}$  are the radial polynomials, i.e.,

$$R_{mn}(\rho) = \sum_{s=0}^{n-|m|} \frac{(-1)^s \rho^{(n-2s)} (n-s)!}{s! (\frac{n+|m|}{2} - s)! (\frac{n-|m|}{2} - s)!}, \quad (5.16)$$

where  $\rho$  and  $\theta$  are the radial coordinates. A ZM of a face image  $I(x, y)$  consisting of a real and an imaginary components is [185]

$$Z_{nm}^I = \frac{n+1}{\pi} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} I(x, y) V_{mn}^*(\rho_{xy}, \theta_{xy}) \Delta \bar{x} \Delta \bar{y}, \quad (5.17)$$

### 5.3. FEATURE EXTRACTION

---

where  $x$  and  $y$  are the image coordinates mapped to the range  $[-1, +1]$ ,  $\rho_{xy} = \sqrt{x^2 + y^2}$ ,  $\theta_{xy} = \tan^{-1} \frac{\bar{x}}{\bar{y}}$  and  $\Delta\bar{x} = \Delta\bar{y} = 2/N\sqrt{2}$ .

Since a local descriptor represents the discontinuities and texture of an image more effectively, QLZM is proposed in [185] using non-linear encoding and pooling, where non-linear encoding facilitates the relevance of low-level features by increasing their robustness against image noise, while pooling is exploited to deal with the problem of small geometric variation. Non-linear encoding is carried out on complex-valued local ZMs using binary quantization, which converts the real and imaginary parts of each ZM coefficient into binary values using signum functions. Such coarse quantisation increases compression and encodes each local block with a single integer. Since features along borders may fall out of the local histogram, they are down-weighted in pooling using a Gaussian window peaked at the centre of each subregion. A second partitioning is also applied to account for the down-weighted features, where a higher emphasis is placed on features down-weighted at the first partitioning. The final QLZM feature is constructed by concatenating all local histograms, and the length of extracted correspond to two parameters: the number of moment coefficient  $K_1$  and the size of the grid  $M$ , which are computed by

$$2^{2K} \times [M^2 + (M + 1)^2], \quad (5.18)$$

where for moment order  $n$ ,  $K_1$  is computed using the function of moment order  $n$

$$K_1(n) = \begin{cases} \frac{n(n+2)}{4} & \text{if } n \text{ is even} \\ \frac{(n+1)^2}{4} & \text{if } n \text{ is odd.} \end{cases} \quad (5.19)$$

The process of generating QLZM is illustrated in Figure 5.5.

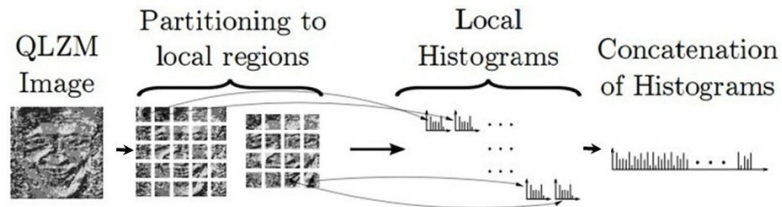


Figure 5.5: QLZM based facial representation framework.

### 5.3.6 Extension to spatio-temporal

QLZM of a 2D image incorporating local spatial textural information has been shown to achieve good facial expression recognition rate [185]. Framework 3 incorporates dynamic information by applying a Motion Change Frequency (MCF) for spatial QLZM, and proposes a spatio-temporal descriptor QLZM\_MCF. Suppose a QLZM sequence where each image frame has been transformed by using QLZM, and the subregions of each QLZM frames are denoted as  $Q_{p,q}(i, t)$ , where  $t$  is the image frame number in the sequence and  $i$  denotes the local pattern from a subregion  $(m, n)$  of each QLZM image. For each pattern  $i$  in subregions  $(p, q)$ , its positive change sequence  $pos_{p,q}(i, t)$ ,  $t = 1, 2, \dots, T - 1$  is defined as

$$pos_{p,q} = \begin{cases} 1 & Q_{p,q}(i, t+1) - Q_{p,q}(i, t) > T_s * Q_{p,q}(i, t) \\ 0 & \text{otherwise} \end{cases} \quad (5.20)$$

where  $T_s$  is a threshold. Similarly, its negative change sequence is defined as

$$neg_{p,q} = \begin{cases} 1 & Q_{p,q}(i, t+1) - Q_{p,q}(i, t) < -T_s * Q_{p,q}(i, t) \\ 0 & \text{otherwise} \end{cases} \quad (5.21)$$

Also, the unchanged sequence is defined as

$$unc_{p,q} = \begin{cases} 1 & |Q_{p,q}(i, t+1) - Q_{p,q}(i, t)| \leq T_s * Q_{p,q}(i, t) \\ 0 & \text{otherwise.} \end{cases} \quad (5.22)$$

$T_s$  is an adjustable parameter which affects the performance of Framework 3. If  $T_s$  is set too large then some movements between two consecutive frames might be ignored, while if  $T_s$  is set too small then small movements, e.g., due to subtle head pose are detected. In our experiments,  $T_s$  is set to 0.1. The QLZM\_MCF on each subregion  $(p, q)$  is the combination of three changes of the pattern  $i$ , i.e.,

$$\begin{aligned} \text{QLZM\_MCF}_{p,q} = \{ & \text{QLZM\_MCF}_{p,q}(i, 1), \\ & \text{QLZM\_MCF}_{p,q}(i, 2), \\ & \text{QLZM\_MCF}_{p,q}(i, 3) \} \end{aligned} \quad (5.23)$$

### 5.3. FEATURE EXTRACTION

---

where

$$\begin{aligned}
\text{QLZM\_MCF}_{p,q}(i, 1) &= \sum_{t=1}^{T-1} \text{pos}(i, t)/(T-1) \\
\text{QLZM\_MCF}_{p,q}(i, 2) &= \sum_{t=1}^{T-1} \text{neg}(i, t)/(T-1) \\
\text{QLZM\_MCF}_{p,q}(i, 3) &= \sum_{t=1}^{T-1} \text{unc}(i, t)/(T-1).
\end{aligned} \tag{5.24}$$

The final QLZM\_MCF feature is obtained by concatenating all QLZM\_MCF<sub>p,q</sub> on each region.

#### 5.3.7 Fusion using weighting function

Given two different types of facial features, an efficient way to combine them is to concatenate the two features to give

$$f_{\text{FUSION}} = (\text{enMHI\_OF}, \text{QLZM\_MCF}), \tag{5.25}$$

where enMHI\_OF and QLZM\_MCF are the two features.

Another combination scheme is also introduced to combine the two features by applying enMHI\_OF feature as weight function in pooling during the generation of QLZM. Specifically, the same strategy of subregion division is used on the input image of MHI and QLZM, and the threshold based on enMHI\_OF is introduced to each subregion of QLZM image to determine which subregions are removed or retained to compute spatial-temporal QLZM\_MCF. If the enMHI\_OF value of a subregion is larger than the threshold, the subregion at the same location in the QLZM image is retained for further processing, otherwise the subregion is removed. The threshold function is defined as

$$R_{p,q} = \begin{cases} R_{p,q} & \text{enMHI\_OF}_{p,q} > T_{en} \\ \text{remove} & \text{otherwise} \end{cases} \tag{5.26}$$

where  $T_{en}$  is the threshold to be set and  $\text{enMHI\_OF}_{p,q}$  is the value of enMHI\_OF in subregion  $(p, q)$ . In our experiments,  $T_{en}$  needs to be set manually in advance. Thus we tested different values by conducting a series of experiments, and 0.8 was finally selected due to the best performance. This scheme is required because subregions with larger enMHI\_OF value indicating more significant motion (thus making larger contribution to recognition) should be allocated larger weights, while subregions



with smaller enMHI\_OF indicating little motion (thus making no or little contribution to recognition) should be allocated smaller weights or removed. The integrated feature is  $f_{\text{WeightedFUSION}}$ , and the dimension of the feature is  $3 \times 2^{2K} \times N_s$ , where  $N_s$  is the number of selected subregions obtained by the thresholding.

## 5.4 Experiments

### 5.4.1 Facial expression datasets

Since the Extended CK dataset (CK+) [94] is widely used for evaluating the performance of facial expression recognition methods and thus facilitates comparison of performances, 327 image sequences from this dataset are again used. Each image sequence from this dataset has various number of frames and starts with the neutral state and ends with the peak phase of a facial expression. Standard leave-one-out cross-validation scheme is used to evaluate the performance of Framework 3 by computing the average recognition rate. One sequence corresponding to an expression is chosen for testing and the remaining sequences of the same expression are used for training. The proposed recognition system was run 327 times on the selected image sequences, and averaged all recognition rates to obtain the final rates.

203 image sequences of MMI [193] are again used, which includes both posed and spontaneous facial expression sequences. All selected sequences are converted into 8-bit grey-scale images with only the sub-sequences from start frame to the frame with the peak expression phase included.

### 5.4.2 Experimental results

The first experiment aims to investigate the effectiveness of the enMHI\_OF feature, and conducted on the CK+ dataset. As the performance of enMHI\_OF might rely on the size of sub-regions and the number of grey levels represented by  $K$ , the experiment is conducted using different sizes and  $K$ . Table 5.1 shows that better performances are achieved using divided grey levels (i.e., using  $K=4, 10, 20$ ) than using the entire 256 grey levels. Also, using sub-regions with size  $20 \times 20$  gives the best performances.

The second experiment compares the performance difference between the spatial and spatio-temporal features. The recognition rates in using spatial QLZM and the spatio-temporal QLZM\_MCF which employs dynamic information are summarised in Figure 5.6. The use of MHI\_OF and MHI is also compared. Since using MHI image as input of classifier may lead to higher dimensionality, histogram computation is used to represent MHI and MHI\_OF. The recognition rates in using MHI

#### 5.4. EXPERIMENTS

Table 5.1: Recognition rate of enMHI\_OF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	$20 \times 20$	$10 \times 10$	$8 \times 8$	$5 \times 5$
K=4	74.31	70.33	71.55	67.28
K=10	75.84	75.84	70.63	72.78
K=20	76.14	75.53	72.48	74.92
K=256	73.40	70.94	71.55	73.09

and MHI\_OF are shown in Figure 5.7. These two figures show that QLZM\_MCF and MHI\_OF outperform the spatial QLZM and MHI, respectively, although the performance of both MHI and MHI\_OF are less than satisfactory.

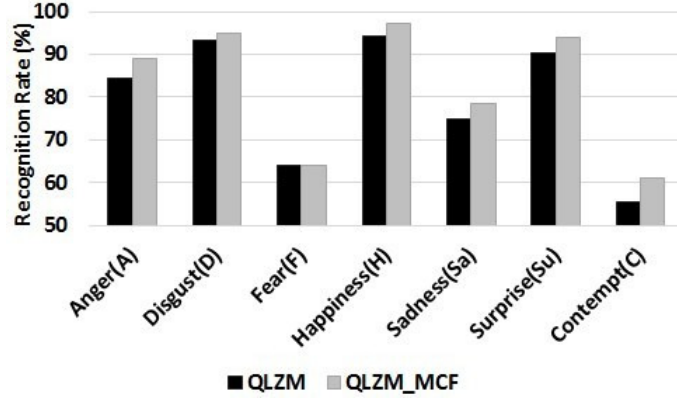


Figure 5.6: Recognition rates of all expressions using QLZM and QLZM\_MCF.

The third experiment investigates the effectiveness of concatenating QLZM\_MCF with enMHI\_OF in Framework 3. Table 5.2, Table 5.3, Table 5.4 and Table 5.5 respectively show the results using two individual features separately, the simple fusion strategy  $f_{\text{FUSION}}$  and the proposed fusion strategy  $f_{\text{WeightedFUSION}}$ . The overall recognition rates using all four features (QLZM\_MCF, enMHI\_OF, simple fused feature and weighting fused feature) are shown in Table 5.6. The tables show the framework using the simple fusion strategy of two features performs better than using individual feature separately, and the proposed fusion strategy achieves the best performance. Table 5.6 shows the comparison in the use of the proposed feature with the method of Eskil et al. [194], the static method of Lucey et al. [189] and Framework 2 [195], which shows the fused feature achieves an average recognition rate of 88.30% for all seven facial expressions, and outperforms the other methods. Thus, it can also be concluded that the combination of two dynamic features improves the recognition rate.

#### 5.4. EXPERIMENTS

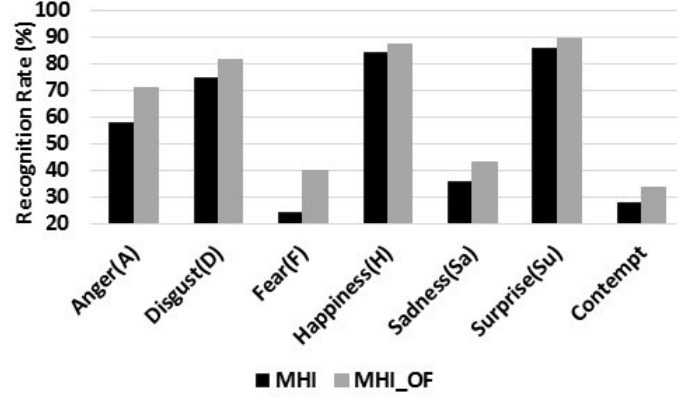


Figure 5.7: Recognition rates of all expressions using MHI and MHI\_OF.

Table 5.2: Recognition rate of QLZM\_MCF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	40	1	0	0	2	0	2
Disgust(D)	1	56	1	0	0	0	1
Fear(F)	1	2	16	2	3	0	1
Happiness(H)	1	0	0	67	0	1	0
Sadness(Sa)	0	1	2	0	22	1	2
Contempt(Su)	0	0	1	2	1	78	1
Contempt(C)	0	1	3	2	1	0	11

Table 5.3: Recognition rate of enMHI\_OF on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	33	2	3	3	3	0	1
Disgust(D)	3	50	2	0	2	0	2
Fear(F)	2	0	10	4	5	1	3
Happiness(H)	0	0	3	62	2	0	2
Sadness(Sa)	3	0	6	1	12	2	4
Surprise(Su)	1	0	2	2	1	75	2
Contempt(C)	2	0	5	1	2	1	7

#### 5.4. EXPERIMENTS

---

Table 5.4: Recognition rate of using simple fusion strategy on classification of six facial expressions and contempt of the CK+ dataset with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	39	1	1	1	3	0	0
Disgust(D)	2	56	1	0	0	0	0
Fear(F)	2	1	18	0	1	1	2
Happiness(H)	0	0	0	66	2	1	0
Sadness(Sa)	1	0	3	0	22	0	2
Surprise(Su)	0	0	1	2	1	79	0
Contempt(C)	1	0	4	1	1	1	10

Table 5.5: Recognition rate of the proposed fusion strategy on classification of six facial expressions of the CK+ dataset and contempt with leave-sequence-out cross-validation.

	A	D	F	H	Sa	Su	C
Anger(A)	41	1	3	0	0	0	0
Disgust(D)	2	57	0	0	0	0	0
Fear(F)	1	1	20	1	0	1	1
Happiness(H)	0	1	0	68	0	0	0
Sadness(Sa)	1	0	0	1	25	1	0
Surprise(Su)	0	0	0	0	1	81	1
Contempt(C)	1	0	2	1	1	0	13

Table 5.6: The overall recognition rates of the four spatio-temporal features on the CK+ dataset.

Feature	Recognition rate
QLZM_MCF	82.6
enMHI_OF	65.7
Eskil et al [194]	76.8
Lucey et al [189]	50.4
Framework 2 [195]	83.7
simple fused feature	82.6
proposed fused feature	88.3

#### 5.4. EXPERIMENTS

---

An experiment was also conducted on the MMI dataset, comparing Framework 3 with the method that uses LBP and SVM [192], [35] and [195] that are evaluated using the same classification strategy of 10-fold cross-validation. The average recognition rates are shown in Table 5.7. The table shows that Framework 3 outperforms all the other five methods. The result for LBP was obtained by using different samples to those used in [192] and using the same strategy of classification introduced in [35] which is also used in [195] and Framework 3.

Table 5.7: Comparative evaluation of Framework 3 on the MMI dataset.

Study	Methodology
LBP [192]	54.5
AAM [35]	62.4
ASM [35]	64.4
Fang in [35]	71.6
Framework 2 [195]	74.3
Framework 3	79.8

Although CK+ and MMI are two of the most widely used datasets for evaluating facial expression recognition methods, they are both collected in a strict controlled settings with near frontal poses, consistent illumination and posed expressions. The recent and more challenging datasets of AFEW [196] and SFEW [197] provide platforms for researchers to create, extend and test their methods on a common benchmarked data. Since Framework 3 recognises facial expression on video sequences that treat a sequence as an entity, image sequences from AFEW which are used for EmotiW 2014 challenge [198] are used for evaluating its performance. AFEW is a dynamic temporal facial expressions data corpus extracted from movies with realistic real world environment. It was collected on the basis of Subtitles for Deaf and Hearing impaired (SDH) and Closed Caption (CC) for searching expression-related content and extracting time stamps corresponding to video clips which represent some meaningful facial motion. The database contains a large age range of subjects from 1-70 years, and the subjects in the clips have been annotated with attributes like Name, Age of Actor, Age of Character, Pose, Gender, Expression of Person and the overall Clip Expression. There are a total of 957 video clips in the database labelled with six basic expressions anger, disgust, fear, happy, sad, surprise and the neutral. Figure 5.8 shows some sample frames from AFEW dataset, and Table 5.8 shows the occurrences of the various expression classes. To compare with the baseline method of EmotiW 2014 [198], Framework 3 is modified slightly, where the pre-processing methods (face detection and alignment) provided by the baseline method are used.

## 5.5. CONCLUSION

---



Figure 5.8: Sample frames from AFEW dataset.

Table 5.8: Number of image sequences (subjects) for each expression in the AFEW dataset.

	<b>A</b>	<b>D</b>	<b>F</b>	<b>H</b>	<b>N</b>	<b>Sa</b>	<b>Su</b>	<b>T</b>
train	118	72	77	145	131	107	73	723
val	64	40	46	63	63	61	46	383

We used the training samples for training, and the validation samples for performance evaluation. Table 5.9 shows the recognition rate using Framework 3 on AFEW dataset. The overall recognition rate of Framework 3 on the validation set is 37.63%, which is higher than the 33.15% achieved by the video only baseline method. Unlike the experiments on CK+ dataset, the surprise expression is much more difficult to be recognised. This is because the surprise expression might not be acted exaggeratedly (i.e., the openness of mouth) sometimes in the real world. Also, the overall recognition rate is much lower than on the CK++ and MMI dataset. This is because numerous frames from the AFEW sequences were captured under poor light condition, have large pose or occlusion, and the expressions are not always from neutral to peak expression.

## 5.5 Conclusion

This chapter presents Framework 3 for facial expression recognition using enMHI\_OF and QLZM\_MCF. The framework which comprises pre-processing, feature extraction followed by 2D PCA and SVM classification achieves better performance than most of the state-of-art methods on CK+ dataset and MMI dataset. The main contributions in Framework 3 are three folds. First, a spatio-temporal feature based on QLZM is proposed. Second, optical flow is applied on MHI to obtain MHI\_OF

## 5.5. CONCLUSION

---

Table 5.9: Recognition rate of the proposed strategy on classification of six basic facial expressions and neutral expression of the AFEW dataset.

	A	D	F	H	N	Sa	Su
Anger(A)	42	5	3	1	6	4	3
Disgust(D)	6	9	3	5	7	4	6
Fear(F)	10	6	9	7	7	4	3
Happiness(H)	2	5	5	40	6	4	1
Neutral(N)	2	4	9	6	31	6	5
Sadness(Sa)	5	7	9	9	16	13	2
Surprise(Su)	6	5	8	6	9	2	10

feature which incorporates velocity information. Third, entropy is introduced to employ the spatial relation of different facial parts, and a strategy is designed based on entropy to integrate enMHL\_OF and QLZM\_MCF. Framework 3 performs slightly worse in distinguishing three expressions of Fear, Sadness and Contempt, thus how to design a better feature to represent these expressions will be part of future work. Also, since an expression usually occurs along with the movement of shoulder and hands, it might be useful to exploit these information in our recognition system.

## Chapter 6

# Conclusions and Future Work

### 6.1 Conclusions

In this thesis, an important field of computer vision has been studied, namely facial expression recognition based on spatio-temporal domain.

Chapter 1 has introduced facial expression recognition and presented the motivation to undertake an investigation on facial expression recognition based on spatio-temporal information. This research interest is motivated by the increasing importance of the role of facial expression recognition on human computer interaction and its wide range of applications in terms of both academic and business. The focus on spatio-temporal information is driven by its advantages in exploiting dynamic facial features. Current challenges in this area have also been presented in this chapter.

Chapter 2 has provided a thorough literature review on techniques of facial expression recognition classified into two categories, i.e., spatio-temporal method and spatial. The spatio-temporal approach analyses a video as a 3D volume using body volumes, interest points or optical flows. It achieves promising performance for simple and periodic facial actions. The spatial approach extracts feature vectors from one image or frame from an image sequence. It is further sub-divided into geometric-based and appearance-based. Geometric-based approach uses the shape and locations of facial components (including mouth, eyes, brows, nose). Appearance-based encodes low-level or high-level information, where the more popular former interprets expressions using low-level histograms and Gabor descriptor, and the latter uses sparse coding [71; 73] or Non-Negative Matrix Factorisation (NMF) [74; 75]. Appearance-based methods are robust to illumination variation and registration errors, while geometric-based methods exploit the movement of facial landmarks and the dimensionality of the feature representation is relatively



low.

In order to obtain discriminative information for analysing facial expressions, most methods extract features based on some mathematical or geometrical heuristics. It has been demonstrated that the task of expression analysis and recognition could be done in more conductive manner by understanding human visual system [72]. Using some salient regions that have significant contribution on recognising facial expression instead of the whole face region might improve the recognition performance. Thus, Chapter 3 has presented Framework 1 based on these salient regions, where some prominent facial patches which rely on the location of facial landmarks are extracted during emotion elicitation. These active patches are then selected and processed to obtain the salient patches which contain discriminative features for classification of each pair of expressions, as different patches have various contributions on different pairs of expression classes. Classifiers using one-against-one strategy are employed to classify these features. An improved sparse representation technique is applied to extract sparse features, which achieves promising performance when compared with state-of-art facial expression recognition methods. This framework replaces feature extraction from the whole face with facial patches, which reduces computational complexity.

A facial expression involves a dynamic process, and the dynamic information such as the movement of facial landmarks and the change in facial shape contains useful information that can represent a facial expression more effectively. Previous recognition methods on video sequences tend to only focus on the movement of facial landmarks, not analysing the variation of facial shape. Chapter 4 has thus presented Framework 2 which utilises two types of dynamic information to enhance the recognition: a novel spatio-temporal descriptor based on PHOG to represent changes in facial shape, and dense optical flow to estimate the movement (displacement) of facial landmarks. The framework views an image sequence as a spatio-temporal volume, and uses temporal information to represent the dynamic movement of facial landmarks associated with a facial expression. In this chapter, PHOG descriptor which represents spatial local shape is extended to spatio-temporal domain so as to capture changes in local shape of facial sub-regions in the temporal dimension to give 3D facial component sub-regions of forehead, mouth, eyebrow and nose that are referred as PHOG\_TOP. By combining PHOG\_TOP and dense optical flow of the facial region, the framework exploits the fusion of discriminant features for classifying and thus recognising facial expressions. A series of experiments were carried out to evaluate the performance of the proposed descriptor, where it is demonstrated that the integrated framework achieves a better performance than using individual

descriptor, and also outperforms most state-of-the-art methods. Framework 2 has been published in [23]

Chapter 5 has proposed a spatio-temporal feature based on local Zernike moment in the spatial domain using motion change frequency. A novel dynamic feature comprising motion history image and entropy is also designed. Entropy-based methods extract intensity information of image pixels. For example, Chai et al. [137] introduced the entropy of a facial region, where a low entropy value means the probabilities of different intensities are different, and a high value means the probabilities are the same. They used the entropy of each of the equal-size blocks of a face image to determine the number of sub-blocks within each block. Inspired by [137], Framework 3 in this chapter applies entropy to the proposed dynamic feature comprising motion history image. Since the intensity value of each pixel in MHI represents a movement, the high intensity values denoting large movement will result in high entropy value, and vice versa. A paper on Framework is currently being considered for publication by the journal Pattern Recognition.

## 6.2 Future work

The performance of the proposed three frameworks is limited by three problems. First, the frameworks have limiting generalisation to other datasets as different datasets are generated under different environments. Second, the frameworks only consider small head pose. When a subject performs a expression along with large head pose, the recognition performance will decrease. Third, the frameworks cannot deal with the large occlusion (e.g., glasses, hair, etc.). Head pose variation remain mostly undressed at representation level. Part based representation or warping the face to the frontal view can address the problem only partially. One possible solution is to learn the relationship between head pose and expression variation at recognition level through statistical modelling . Another is to address head pose variation via high-level representation that learns the appearance variation caused by head pose variations using linear or non-linear feature extraction techniques such as factor analysis. Also, high-level representations are highly promising for dealing with identity bias.

Although high-level representations are promising for dealing with identity bias and head pose variation, they are not been exploited to their full potential. One future direction is to develop efficient shape representation paradigms. The shift towards appearance-based representations is mainly due to the registration sensitivity of shape representations. However, registration sensitivity is an issue of existing representations rather than shape-based representation in general. Shape

representations deserve attention for multiple reasons. From a cognitive science perspective, they can be argued to play an important role in human vision for the perception of facial expressions [199]. From a computer vision perspective, they are invariant to illumination variations and less sensitive to identity bias than appearance representations. Shape representations can describe continuous shape rather than discrete points (e.g. [200], [201]).

One way to deal with head-pose variations is to design high-level representations that learn the appearance variation caused by head-pose variations using linear or nonlinear feature extraction techniques such as factor analysis [202], multi-linear mapping with tensors [203] or manifold learning [148]. However, the amount of texture variation induced by head-pose variations can be too difficult to handle even for such sophisticated methods [204]. Developing high-level part-based representations is an approach that can prove to be more successful in other domains such as face recognition [202]. Once the spatial consistency of spatially distant parts is ensured through parts registration, modelling the within-part appearance variation can potentially be simpler with high-level representations. Overall, high-level representations can play an important role in affect recognition, but their design requires a special care.

Like most facial expression recognisers, the three frameworks still rely on 2D images as input. The rapid progress in depth-based imaging technology is supporting 3D face analysis by overcoming the challenges associated to head-pose and illumination variations. Moreover, the analysis of depth variations facilitates the recognition of expressions that might be hardly noticeable using only 2D appearance [205].

3D facial expression analysis methods share conceptual similarities with those based on 2D analysis. 3D expression analysers often perform registration using techniques that, similar to the registration techniques discussed, model a face as a collection of facial landmarks [206], [207]. 3D shape representations involve features such as the angles and distances between landmarks [208], which resemble the features of 2D shape representations. In several cases, 3D data are projected onto 2D images, which are then represented using well-established 2D representations, such as Gabor [207] or SIFT-based approaches [97]. Dimensionality reduction techniques, such as Boosting-based methods [207] or [209], or statistical models, such as SVM [207] or HMM [210], are commonly employed in 2D and 3D analyses.

The future research in facial expression recognition involves: designing an algorithm based on high-level representation that is robust against illumination changes and occlusions; investigating 3D facial expression analysis by adding depth

## 6.2. *FUTURE WORK*

---

information; improving the generalisation of the framework.

# Bibliography

- [1] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.
- [2] “Universals and cultural differences in the judgments of facial expressions of emotion.,” *Ekman, Paul and Friesen, Wallace V and O’Sullivan, Maureen and Chan, Anthony and Diacoyanni-Tarlatzis, Irene and Heider, Karl and Krause, Rainer and LeCompte, William Ayhan and Pitcairn, Tom and Ricci-Bitti, Pio E and others*, vol. 53, no. 4, p. 712, 1987.
- [3] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton Company, 2009.
- [4] A. parametric model for human faces., “A parametric model for human faces.,” tech. rep., DTIC Document, 1974.
- [5] J. M. Carroll and J. A. Russell, “Do facial expressions signal specific emotions? judging emotion from the face in context,” *Journal of personality and social psychology*, vol. 70, no. 2, p. 205, 1996.
- [6] J. A. Russell, “Culture and the categorization of emotions.,” *Psychological bulletin*, vol. 110, no. 3, p. 426, 1991.
- [7] J. A. Russell, “Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies.,” *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.
- [8] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.
- [9] P. Ekman and W. V. Friesen, “Unmasking the face: A guide to recognizing emotions from facial cues.” Englewood Cliffs, NJ: Prentice Hall, 1975.

- [10] P. Ekman, “Facial expression and emotion.,” *American psychologist*, 1993.
- [11] P. Bull, “State of the art: Nonverbal communication,” *The Psychologist*, vol. 14, no. 12, pp. 644–647, 2001.
- [12] V. H. Yngve, “On getting a word in edgewise,” in *Chicago Linguistics Society, 6th Meeting*, pp. 567–578, 1970.
- [13] P. Carrera-Levillain and J.-M. Fernandez-Dols, “Neutral faces in context: Their emotional meaning and their function,” *Journal of Nonverbal Behavior*, 1994.
- [14] J.-M. Fernández-Dols, H. Wallbott, and F. Sanchez, “Emotion category accessibility and the decoding of emotion from facial expression and context,” *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 107–123, 1991.
- [15] N. Ambady and R. Rosenthal, “Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis,” *Psychological bulletin*, 1992.
- [16] J. Cohn and P. Ekman, “Measuring facial action by manual coding, facial emg, and automatic facial image analysis,” in *Handbook of nonverbal behavior research methods in the affective sciences*, Carnegie Mellon University, 2003.
- [17] C. E. Izard, “Emotion theory and research: Highlights, unanswered questions, and emerging issues,” *Annual review of psychology*, vol. 60, p. 1, 2009.
- [18] P. Ekman and W. Freisen, “Facial action coding system (facs): A technique for the measurement of facial expression. paolo alto,” 1978.
- [19] Y.-l. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.
- [20] M. F. Valstar, *Timing is everything: A spatio-temporal approach to the analysis of facial actions*. PhD thesis, Imperial College London, 2008.
- [21] B. Fasel, F. Monay, and D. Gatica-Perez, “Latent semantic analysis of facial action codes for automatic facial expression recognition,” in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 181–188, 2004.
- [22] G. C. Littlewort, M. S. Bartlett, and K. Lee, “Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain,”

in *In Proceedings of the International Conference on Multimodal Interfaces*, pp. 15–21, 2007.

- [23] D. Chetverikov and R. Péteri, “A brief survey of dynamic texture description and recognition,” in *Computer Recognition Systems*, pp. 17–26, Springer, 2005.
- [24] M. Pantic and I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 2, pp. 433–449, 2006.
- [25] K. L. Schmidt, J. F. Cohn, and Y. Tian, “Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles,” *Biological psychology*, vol. 65, no. 1, pp. 49–66, 2003.
- [26] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre, “Detecting depression from facial actions and vocal prosody,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACHI 2009. 3rd International Conference on*, pp. 1–7, 2009.
- [27] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, *et al.*, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.
- [28] B. Heisele, P. Ho, J. Wu, and T. Poggio, “Face recognition: component-based versus global approaches,” *Computer vision and image understanding*, vol. 91, no. 1, pp. 6–21, 2003.
- [29] C. Singh, N. Mittal, and E. Walia, “Face recognition using zernike and complex zernike moment features,” *Pattern Recognition and Image Analysis*, vol. 21, no. 1, pp. 71–81, 2011.
- [30] G. Yang and T. S. Huang, “Human face detection in a complex background,” *Pattern Recognition*, vol. 27, no. 1, pp. 53 – 63, 1994.
- [31] A. V. Nefian and M. H. Hayes III, “Face detection and recognition using hidden markov models,” in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, vol. 1, pp. 141–145, 1998.
- [32] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod, “Model-based face tracking for view-independent facial expression recognition,” in *Automatic*

*Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 287–293, 2002.

- [33] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod, “Model-based face tracking for view-independent facial expression recognition,” in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 287–293, IEEE, 2002.
- [34] M. Pantic and L. J. M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.
- [35] H. Fang, N. Mac Parthaláin, A. J. Aubrey, G. K. Tam, R. Borgo, P. L. Rosin, P. W. Grant, D. Marshall, and M. Chen, “Facial expression recognition in dynamic sequences: An integrated approach,” *Pattern Recognition*, vol. 47, no. 3, pp. 1271–1281, 2014.
- [36] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou, “Image ratio features for facial expression recognition application,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 3, pp. 779–788, 2010.
- [37] K. K. Sung and T. Poggio, “Example-based learning for view-based human face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 39–51, Jan 1998.
- [38] M.-H. Yang, D. Kriegman, and N. Ahuja, “Detecting faces in images: a survey,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.
- [39] C. Kotropoulos and I. Pitas, “Rule-based face detection in frontal views,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 4, pp. 2537–2540, 1997.
- [40] H. Mekami and S. Benabderrahmane, “Towards a new approach for real time face detection and normalization,” in *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, pp. 445–459, 2010.
- [41] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A survey of skin-color modeling and detection methods,” *Pattern Recogn.*, vol. 40, pp. 1106–1122, Mar. 2007.



- [42] S. L. Phung, A. Bouzerdoun, and D. Chai, "Skin segmentation using color pixel classification: analysis and comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 148–154, Jan 2005.
- [43] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 696–706, 2002.
- [44] M. Kawulok, J. Kawulok, and J. Nalepa, "Spatial-based skin detection using discriminative skin-presence features," *Pattern Recognition Letters*, vol. 41, pp. 3–13, 2014.
- [45] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell, "A fusion approach for efficient human skin detection," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 138–147, 2012.
- [46] M. F. Augusteijn and T. L. Skufca, "Identification of human faces through texture-based feature recognition and neural network technology," in *Neural Networks, 1993., IEEE International Conference on*, pp. 392–398, IEEE, 1993.
- [47] S. Fahlman and C. Lebiere, "The cascade-correlation learning architecture; advances in neural information systems," 1990.
- [48] T. Kohonen, "Self-organising and associative memory," *Springer Series on Information Sciences (Springer-Verlag)*, 1989.
- [49] P. Sinha, "Perceiving and recognizing three-dimensional forms /," *Massachusetts Institute of Technology*, 1995.
- [50] B. Scassellati, "Eye finding via face detection for a foveated, active vision system," in *Fifteenth National/tenth Conference on Artificial Intelligence/innovative Applications of Artificial Intelligence*, p. 189201, 1998.
- [51] K. Anderson and P. W. Mcowan, "Robust real-time face tracker for cluttered environments," *Computer Vision & Image Understanding*, vol. 95, no. 2, pp. 184–200, 2004.
- [52] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [53] H. Rowley, S. Baluja, T. Kanade, *et al.*, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23–38, 1998.

- [54] E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: an application to face detection,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997.
- [55] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37, 1995.
- [56] C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in *Computer vision, 1998. sixth international conference on*, pp. 555–562, 1998.
- [57] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001.
- [58] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, “Robust continuous prediction of human emotions using multiscale dynamic cues,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 501–508, 2012.
- [59] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang, “Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior,” *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 79–91, 2011.
- [60] L. Zhang and D. Tjondronegoro, “Facial expression recognition using facial movement features,” *Affective Computing, IEEE Transactions on*.
- [61] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, “Facial point detection using boosted regression and graph models,” vol. 119, no. 5, pp. 2729–2736, 2010.
- [62] D. Vukadinovic and M. Pantic, “Fully automatic facial feature point detection using gabor feature based boosted classifiers,” in *IEEE International Conference on Systems, Man and Cybernetics*, pp. 1692–1698, 2005.
- [63] I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for tracking facial features,” in *IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings*, p. 97, 2004.

- [64] X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognition*, 2015.
- [65] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 6, pp. 1113–1133, 2014.
- [66] S. Lucey, A. B. Ashraf, and J. F. Cohn, *Investigating spontaneous facial action recognition through aam representations of the face*. INTECH Open Access Publisher, 2007.
- [67] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.
- [68] K.-C. Huang, S.-Y. Huang, and Y.-H. Kuo, "Emotion recognition based on a novel triangular facial feature extraction method," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–6, 2010.
- [69] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1449–1461, 2004.
- [70] T. Senechal, V. Rapp, and L. Prevost, "Facial feature tracking for emotional dynamic analysis," in *Advanced Concepts for Intelligent Vision Systems*, pp. 495–506, 2011.
- [71] S. F. Cotter, "Sparse representation for accurate classification of corrupted and occluded facial expressions," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 838–841, 2010.
- [72] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 336–342, 2011.
- [73]

- [74] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, “Subclass discriminant non-negative matrix factorization for facial image analysis,” *Pattern Recognition*, vol. 45, no. 12, pp. 4080–4091, 2012.
- [75] R. Zhi, M. Flierl, Q. Ruan, and B. W. Kleijn, “Graph-preserving sparse non-negative matrix factorization with application to facial expression recognition,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 1, pp. 38–52, 2011.
- [76] L. Jeni, J. M. Girard, J. F. Cohn, F. De La Torre, *et al.*, “Continuous au intensity estimation using localized, sparse facial feature space,” in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–7, 2013.
- [77] G. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett, “The motion in emotion—a cert based approach to the fera emotion challenge,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 897–902, 2011.
- [78] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan, “Action unit recognition transfer across datasets,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 889–896, 2011.
- [79] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, *et al.*, “Multiple classifier systems for the classification of audio-visual emotional states,” in *Affective Computing and Intelligent Interaction*, pp. 359–368, Springer, 2011.
- [80] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. V. d Malsburg, R. Wurtz, and W. Konen, “Distortion invariant object recognition in the dynamic link architecture,” *Computers, IEEE Transactions on*, vol. 42, no. 3, pp. 300–311, 1993.
- [81] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von Der Malsburg, “Face recognition by elastic bunch graph matching,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pp. 775–779, 1997.
- [82] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, “Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 202–207, 2000.

- [83] Y.-l. Tian, T. Kanade, and J. F. Cohn, “Eye-state action unit detection by gabor wavelets,” in *Advances in Multimodal Interfaces—ICMI 2000*, pp. 143–150, Springer, 2000.
- [84] Y.-l. Tian, T. Kanade, and J. F. Cohn, “Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity,” in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 229–234, 2002.
- [85] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Classifying facial actions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 974–989, 1999.
- [86] G. Littlewort-Ford, M. S. Bartlett, and J. R. Movellan, “Are your eyes smiling? detecting genuine smiles with support vector machines and gabor wavelets,” in *Proceedings of the 8th Joint Symposium on Neural Computation*, 2001.
- [87] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” in *Computer vision-eccv 2004*, Springer, 2004.
- [88] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.
- [89] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [90] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, “Emotion recognition using phog and lpq features,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 878–883, 2011.
- [91] A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel,” in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 401–408, ACM, 2007.
- [92] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

- [93] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 4, no. 7, pp. 971–987, 2002.
- [94] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101, 2010.
- [95] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, 2005.
- [96] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598–1603, 2011.
- [97] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.
- [98] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.
- [99] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 401–408, ACM, 2007.
- [100] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, 2005.
- [101] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, pp. 2169–2178, 2006.

- [102] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro, “Local zernike moment representations for facial affect recognition,” in *Proc. of British Machine Vision Conf*, 2013.
- [103] M.-K. Hu, “Visual pattern recognition by moment invariants,” *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, 1962.
- [104] M. R. Teague, “Image analysis via the general theory of moments,” *JOSA*, vol. 70, no. 8, pp. 920–930, 1980.
- [105] A. Khotanzad and Y. H. Hong, “Invariant image recognition by zernike moments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 489–497, May 1990.
- [106] A. Ono, “Face recognition with zernike moments,” *Systems and Computers in Japan*, 2003.
- [107] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, no. 1, pp. 1457–1469, 2004.
- [108] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.
- [109] M. W. Huang, Z. W. Wang, and Z. L. Ying, “A new method for facial expression recognition based on sparse representation plus lbp,” in *International Congress on Image and Signal Processing*, pp. 6826–6829, 2010.
- [110] S. Zhang, X. Zhao, and B. Lei, “Robust facial expression recognition via compressive sensing,” *Sensors*, vol. 12, no. 3, pp. 3747–61, 2012.
- [111] Z. W. Wang, M. W. Huang, and Z. L. Ying, “The performance study of facial expression recognition via sparse representation,” in *International Conference on Machine Learning and Cybernetics, ICMLC 2010, Qingdao, China, July 11-14, 2010, Proceedings*, pp. 824–827, 2010.
- [112] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 2037–2041, Dec 2006.
- [113] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, “Authentic facial expression analysis,” *Image & Vision Computing*, vol. 25, no. 12, pp. 1856–1863, 2004.

- [114] S. M. Lajevardi and Z. M. Hussain, “Automatic facial expression recognition: feature extraction and selection,” *Signal, Image and video processing*, vol. 6, no. 1, pp. 159–169, 2012.
- [115] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.
- [116] T. Jabid, M. H. Kabir, and O. Chae, “Robust facial expression recognition based on local directional pattern,” *ETRI journal*, vol. 32, no. 5, pp. 784–794, 2010.
- [117] H. Kabir, T. Jabid, and O. Chae, “A local directional pattern variance (ldpv) based face descriptor for human facial expression recognition,” in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 526–532, IEEE, 2010.
- [118] C. Shan and R. Braspenning, “Recognizing facial expressions automatically from video,” in *Handbook of ambient intelligence and smart environments*, pp. 479–509, Springer, 2010.
- [119] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, “Learning active facial patches for expression analysis,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2562–2569, IEEE, 2012.
- [120] C. Shan, S. Gong, and P. W. McOwan, “Robust facial expression recognition using local binary patterns,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2, pp. II–370, IEEE, 2005.
- [121] C. Shan and T. Gritti, “Learning discriminative lbp-histogram bins for facial expression recognition,” in *BMVC*, pp. 1–10, 2008.
- [122] S. Moore and R. Bowden, “Local binary patterns for multi-view facial expression recognition,” *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011.
- [123] I. Kotsia, S. Zafeiriou, and I. Pitas, “Texture and shape information fusion for facial expression and facial action unit recognition,” *Pattern Recognition*, vol. 41, no. 3, pp. 833–851, 2008.
- [124] T. Bnziger and K. R. Scherer, “Introducing the geneva multimodal emotion portrayal (gemep) corpus,” *A Blueprint for Affective Computing A Sourcebook and Manual*, 2010.



- [125] G. Zhao and M. Pietikäinen, “Boosted multi-resolution spatiotemporal descriptors for facial expression recognition,” *Pattern recognition letters*, vol. 30, no. 12, pp. 1117–1127, 2009.
- [126] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, “A dynamic appearance descriptor approach to facial actions temporal modeling,” *Cybernetics, IEEE Transactions on*, vol. 44, no. 2, pp. 161–174, 2014.
- [127] B. Jiang, M. F. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 314–321, 2011.
- [128] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, “Learning spatiotemporal features by using independent component analysis with application to facial expression recognition,” *Neurocomputing*, vol. 93, pp. 126–132, 2012.
- [129] T. Wu, M. S. Bartlett, and J. R. Movellan, “Facial expression recognition using gabor motion energy filters,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 42–47, 2010.
- [130] M. F. Valstar and M. Pantic, “Combined support vector machines and hidden markov models for modeling facial action temporal dynamics,” in *Human-Computer Interaction*, pp. 118–127, Springer, 2007.
- [131] M. Kenji, “Recognition of facial expression from optical flow,” *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.
- [132] S. Koelstra, M. Pantic, and I. Y. Patras, “A dynamic texture-based approach to recognition of facial actions and their temporal models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 11, pp. 1940–1954, 2010.
- [133] J. W. Davis and A. E. Bobick, “The representation and recognition of human movement using temporal templates,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 928–934, 1997.
- [134] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, “Nonrigid registration using free-form deformations: application to

- breast mr images,” *Medical Imaging, IEEE Transactions on*, vol. 18, no. 8, pp. 712–721, 1999.
- [135] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.
  - [136] L. A. Cament, L. E. Castillo, J. P. Perez, F. J. Galdames, and C. A. Perez, “Fusion of local normalization and gabor entropy weighted features for face identification,” *Pattern Recognition*, vol. 47, no. 2, pp. 568–577, 2014.
  - [137] Z. Chai, H. Mendez-Vazquez, R. He, Z. Sun, and T. Tan, “Semantic pixel sets based local binary patterns for face recognition,” in *Computer Vision–ACCV 2012*, Springer, 2013.
  - [138] M. Pantic and L. J. Rothkrantz, “Expert system for automatic analysis of facial expressions,” *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.
  - [139] L. Ma and K. Khorasani, “Facial expression recognition using constructive feedforward neural networks,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1588–1595, 2004.
  - [140] L. Rabiner and B.-H. Juang, “Fundamentals of speech recognition,” 1993.
  - [141] F. S. Samaria, *Face recognition using hidden Markov models*. PhD thesis, University of Cambridge, 1994.
  - [142] M. S. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. J. Sejnowski, and J. R. Movellan, “Automatic analysis of spontaneous facial behavior: A final project report,” tech. rep., Technical Report UCSD MPLab TR 2001.08, University of California, San Diego, 2001.
  - [143] G. Littlewort, I. Fasel, M. S. Bartlett, and J. R. Movellan, “Fully automatic coding of basic expressions from video,” *University of California, San Diego, San Diego, CA*, vol. 92093, 2002.
  - [144] I. Essa, A. P. Pentland, *et al.*, “Coding, analysis, interpretation, and recognition of facial expressions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 757–763, 1997.
  - [145] N. Gueorguieva, G. Georgiev, and I. Valova, “Facial expression recognition using feedforward neural networks,” in *IC-AI*, pp. 285–291, 2003.

- [146] S. Koelstra and M. Pantic, “Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on*, pp. 1–8, 2008.
- [147] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.
- [148] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 46–53, 2000.
- [149] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 5–pp, IEEE, 2005.
- [150] A. Dhall *et al.*, “Collecting large, richly annotated facial-expression databases from movies,” *MultiMedia, IEEE*, vol. 19, no. 3, pp. 34–41, 2012.
- [151] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark,” in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 2106–2112, 2011.
- [152] T. Bänziger, M. Mortillaro, and K. R. Scherer, “Introducing the geneva multimodal expression corpus for experimental research on emotion perception,” *Emotion*, vol. 12, no. 5, p. 1161, 2012.
- [153] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, “Disfa: A spontaneous facial action intensity database,” *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 151–160, 2013.
- [154] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, “Automatic recognition of facial actions in spontaneous expressions,” *Journal of multimedia*, vol. 1, no. 6, pp. 22–35, 2006.
- [155] H. Gunes and M. Piccardi, “A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 1148–1153, 2006.

- [156] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 64–84, 2009.
- [157] K. R. Scherer and H. Ellgring, "Multimodal expression of emotion: Affect programs or componential appraisal patterns?," *Emotion*, vol. 7, no. 1, p. 158, 2007.
- [158] C. A. Smith, G. J. McHugo, and J. T. Lanzetta, "The facial muscle patterning of posed and imagery-induced expressions of emotion by expressive and nonexpressive posers," *Motivation and Emotion*, vol. 10, no. 2, pp. 133–157, 1986.
- [159] M. Hoque, L.-P. Morency, and R. W. Picard, "Are you friendly or just polite?—analysis of smiles in spontaneous face-to-face interactions," in *Affective Computing and Intelligent Interaction*, Springer, 2011.
- [160] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *Proceedings of the 8th international conference on Multimodal interfaces*, pp. 162–170, 2006.
- [161] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 02, pp. 121–132, 2004.
- [162] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pp. 881–888, 2013.
- [163] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 2106–2111, 2009.
- [164] R. C. Schultz, "Anthropometric facial proportions in medicine," *Jama the Journal of the American Medical Association*, vol. 258, no. 9, p. 1245, 1987.
- [165] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual face coding," *Psychophysiology*, vol. 36, no. 01, pp. 35–43, 1999.

- [166] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman, "Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 491, 2002.
- [167] W. E. Rinn, "The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions.," *Psychological Bulletin*, vol. 95, no. 1, pp. 52–77, 1984.
- [168] J. M. Vanswearingen, J. F. Cohn, and A. Bajaj-Luthra, "Specific impairment of smiling increases the severity of depressive symptoms in patients with facial neuromuscular disorders.," *Aesthetic Plastic Surgery*, vol. 23, no. 6, pp. 416–423, 1999.
- [169] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [170] L. Unzueta, W. Pimenta, J. Goenetxea, L. P. Santos, and F. Dornaika, "Efficient generic face model fitting to images and videos," *Image and Vision Computing*, vol. 32, no. 5, pp. 321–334, 2014.
- [171] D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami, "Real-time face detection and lip feature extraction using field-programmable gate arrays," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 4, pp. 902–912, 2006.
- [172] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2002.
- [173] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.
- [174] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 1, pp. 103–108, 1990.
- [175] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.
- [176] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The japanese female facial expression (jaffe) database," 1998.

- [177] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” in *IJCAI*, vol. 81, pp. 674–679, 1981.
- [178] G. L. M.S. Bartlett, I. F. M. Frank, C. Lainscsek, and J. Movellan, “Recognizing facial expression: machine learning and application to spontaneous behavior,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 568–573, June 2005.
- [179] T. K. Ying-Li Tian and J. F. Cohn, “Recognizing action units for facial expression analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.
- [180] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic, “Local evidence aggregation for regression-based facial point detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 5, pp. 1149–1163, 2013.
- [181] Z. Li, J.-i. Imai, and M. Kaneko, “Facial-component-based bag of words and phog descriptor for facial expression recognition,” in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pp. 1353–1358, IEEE, 2009.
- [182] M. T. Eskin and K. S. Benli, “Facial expression recognition based on anatomy,” *Computer Vision and Image Understanding*, vol. 119, pp. 1–14, 2014.
- [183] A. Ono, “Face recognition with zernike moments,” *Systems and Computers in Japan*, vol. 34, no. 10, pp. 26–35, 2003.
- [184] N. M. C. Singh and E. Walia, “Face recognition with zernike moments,” *Pattern Recognition and Image Analysis*, vol. 21, pp. 71–81, 2011.
- [185] *19th IEEE International Conference on Image Processing, ICIP 2012, Lake Buena Vista, Orlando, FL, USA, September 30 - October 3, 2012*, IEEE, 2012.
- [186] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.
- [187] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic, “Local evidence aggregation for regression-based facial point detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 5, pp. 1149–1163, 2013.

- [188] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, “Two-dimensional pca: a new approach to appearance-based face representation and recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.
- [189] P. Lucey, “The extended cohn-kande dataset (ck+): A complete facial expression dataset for action unit and emotion-specified expression. human communicative behavior analysis,” in *Workshop of CVPR*, 2010.
- [190] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” in *IJCAI*, vol. 81, pp. 674–679, 1981.
- [191] R. Balian, “Entropy, a protean concept,” in *Poincaré Seminar 2003*, pp. 119–144, Springer, 2004.
- [192] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [193] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 5–pp, IEEE, 2005.
- [194] M. T. Eskil and K. S. Benli, “Facial expression recognition based on anatomy,” *Computer Vision and Image Understanding*, vol. 119, pp. 1–14, 2014.
- [195] X. Fan and T. Tjahjadi, “A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences,” *Pattern Recognition*, vol. 48, no. 11, pp. 3407–3416, 2015.
- [196] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Collecting large, richly annotated facial-expression databases from movies,” *IEEE MultiMedia*, vol. 19, pp. 34–41, July 2012.
- [197] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark,” in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 2106–2112, Nov 2011.
- [198] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, “Emotion recognition in the wild challenge 2014: Baseline, data and protocol,” in *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI ’14*, (New York, NY, USA), pp. 461–466, ACM, 2014.

- [199] A. M. Martinez, “Deciphering the face,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pp. 7–12, IEEE, 2011.
- [200] S. Lee, “Symmetry-driven shape description for image retrieval,” *Image and Vision Computing*, vol. 31, no. 4, pp. 357–363, 2013.
- [201] D. Zhang and G. Lu, “Review of shape representation and description techniques,” *Pattern recognition*, vol. 37, no. 1, pp. 1–19, 2004.
- [202] S. J. Prince, J. Warrell, J. H. Elder, and F. M. Felisberti, “Tied factor analysis for face recognition across large pose differences,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 6, pp. 970–984, 2008.
- [203] M. A. O. Vasilescu and D. Terzopoulos, “Multilinear analysis of image ensembles: Tensorfaces,” in *Computer Vision ECCV 2002*, pp. 447–460, Springer, 2002.
- [204] B. Tunç, V. Dağlı, and M. Gökmen, “Class dependent factor analysis and its application to face recognition,” *Pattern Recognition*, vol. 45, no. 12, pp. 4092–4102, 2012.
- [205] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, “Static and dynamic 3d facial expression recognition: A comprehensive survey,” *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.
- [206] M. Kaiser, B. Kwolek, C. Staub, and G. Rigoll, “Registration of 3d facial surfaces using covariance matrix pyramids,” in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 1002–1007, IEEE, 2010.
- [207] A. Savran, B. Sankur, and M. T. Bilge, “Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units,” *Pattern recognition*, vol. 45, no. 2, pp. 767–782, 2012.
- [208] H. Tang and T. S. Huang, “3d facial expression recognition based on properties of line segments connecting facial feature points,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on*, pp. 1–6, IEEE, 2008.
- [209] T. Sha, M. Song, J. Bu, C. Chen, and D. Tao, “Feature level analysis for 3d facial expression recognition,” *Neurocomputing*, vol. 74, no. 12, pp. 2135–2141, 2011.



- [210] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, “Recognition of 3d facial expression dynamics,” *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.